AFOSR/SDC

PROCEEDINGS OF THE WORKSHOP ON

Working with Semi-Automatic Documentation Systems

AIRLIE FOUNDATION—WARRENTON, VIRGINIA

This document was produced by System Development Corporation in performance of Contract AF 19(628)-3418, AFOSR Workshop, for the Air Force Office of Scientific Research.

Preface

These proceedings follow the order of the actual conference program. Each of the three papers making up the first part describes a major document system; the seven reports of the second part summarize the panel discussions that occupied most of the three days of the conference. The original program had scheduled eight panels, but it was decided at the outset of the conference to merge input processing with request processing and deal with both in a joint discussion. With the exception of the keynote address by Mr. William Knox, the papers and reports are presented here substantially intact. His speech, however, has been excerpted from his original notes and his extemporaneous remarks. The version attributed to him here is largely the editor's reconstruction.

This report is published by the System Development Corporation and is available to qualified requesters through the Defense Documentation Center, Cameron Station, Alexandria, Virginia 22314.

J. J. MAHER, Editor
System Development Corporation

Workshop Organization

Sponsors

U. S. Air Force Office of Scientific Research
System Development Corporation

Planning Committee

Harold Wooster, U. S. Air Force Office of Scientific Research, *Co-chairman*
Raymond P. Barrett, System Development Corporation, *Co-chairman*
Joseph J. Breen, System Development Corporation
Captain Thomas K. Burgess, U. S. Air Force Office of Scientific Research
Wallace G. Patton, System Development Corporation
Hans C. Ullmann, System Development Corporation

Speakers

William T. Knox, White House Office of Science and Technology
Raymond P. Barrett, System Development Corporation
C. Allen Merritt, International Business Machines Corporation
Audrey S. Williams, Douglas Aircraft Company, Inc.

Panel Chairmen

William Barden, Defense Documentation Center
Bernard K. Dennis, Battelle Memorial Institute
William Hammond, Datatrol Corporation
Herbert Rehbock, Defense Documentation Center
James W. Singleton, System Development Corporation
Y. S. Touloukian, Purdue University
Van A. Wente, National Aeronautics and Space Administration
Fred H. Wise, System Development Corporation

Contents

Raymond P. Barrett
Systems Development Corporation
Workshop Co-chairman

Introduction

We present these proceedings with mixed feelings in that it seems paradoxical for us to comunicate the results of the Workshop by one of the most difficult-to-handle documents to be found in any information system. Our conviction, nevertheless, is strong that these proceedings provide an accurate description of what is being done by applications people today.

The Government and several private organizations are spending large sums of money to support the research and development of semiautomatic systems for handling technical data and documents. Yet, little information of directly usable value has become available to the implementers of information systems. This is understandable, since in our field, as in other fields, research usually precedes implementation by a significant period of time. Researchers, it seems, advance their knowledge through dialogues with other researchers. Applications people, on the contrary, are often too pressed by here-and-now problems; so, they usually find their solutions by pragmatic action rather than by abstract discussion.

For some time prior to this Workshop, the Air Force Office of Scientific Research and the System Development Corporation had been concerned about the unsatisfied information needs of applications people. Each organization was aware that the applications people themselves were the source of the most valuable information. But the chief difficulty in exploiting this source lay in finding some way of bringing applications people together for an exchange of ideas and experiences. Being engaged in the implementation and operation of information systems, applications people gear themselves to the practical demands of daily work. Most do not make it a practice to report on their work to the world at large or to frequent professional meetings. Their common feeling seems to be that the amount of immediately usable information gained by attending meetings is often not worth the required investment of

valuable time. Taking these things into consideration, AFOSR and SDC nevertheless decided that a carefully designed conference could help in bridging the information gap between research and implementation. This Workshop was conceived to carry out this decision.

When representatives of AFOSR and SDC met to plan the Workshop, we found mutual reluctance to sponsor another forum for airing potentially good but remotely usable ideas. We believed that such a forum would not let us accomplish our objectives and most certainly would not attract the desired audience. The planning committee therefore drafted the following points:

1. Attendance limited by invitation to no more than 100 participants actively concerned with information systems applications.
2. Limitation on the reading of papers to let participants spend maximum time in panel discussions.
3. Selection of panel chairmen with a good deal of practical experience in the subject matter of their respective panel discussions.
4. Focusing the Workshop on document systems rather than on specialized systems for management, command-control, and the like.
5. Publication of the speeches and summaries of the panel discussions for the benefit of nonparticipants.

The planning committee worked out a list of topics, recommended the systems to be formally described, and drafted a list of candidates for chairing the panel discussions. The committee also recommended a limit of no more than three speeches plus a keynote address, so that the main business of the Workshop could be concentrated in seven panel discussions covering basic areas of information systems applications work. Finally, the committee recommended that SDC administer the Workshop and assume responsibility for publishing the proceedings.

We did not expect that the Workshop would solve any of the pressing problems that are encountered in the implementation and operation of information systems. But we did expect an accurate picture of where the applications business stands today. In this we were not disappointed.

Harold Wooster
Air Force Office of Scientific Research
Workshop Co-chairman

Summary

There is a story of the eager young County Agent trying to inveigle a hard-bitten dirt farmer into taking a short course at State: "If you'll only go there and listen to all the latest discoveries and look at all the wonderful things they're doing on their test plots, why you'll be able to run your farm twice as efficiently as you're doing now."

"Young feller," said the farmer, "I'm only farming now half as good as I know how to."

In the world of documentation it is almost impossible to avoid listening to the professors and the lightning rod salesmen—the farmers are not heard because they are too busy to talk. This Workshop was a meeting for documentalists who normally don't go to meetings: they have work to do, even if they're doing it now only half as good as they know how to. Here, then, is a summary of what this group of documentalists had to say.

The *indexing and classification* people were reconciled to using some sort of manual coordinate indexing, with later computer juggling of terms. They wondered, though, what they are supposed to do with the twenty-two authorized COSATI subject groups that don't seem to fit the "real" user's world.

If people weren't around to do *abstracting and extracting*, it would be necessary to invent them. People are easier than machines to reprogram to produce the various kinds of abstracts needed.

Vocabulary construction and control panelists agreed there is a need for some sort of subject indexing authority having a display of generic relationships between terms and instructions to indexers for using the terms. The computer makes a useful clerk-editor-monitor for keeping track of what the people are doing. Serial vs. inverted files remains a moot question, but serial files gain a slight advantage where computer capacity permits their use.

Character readers aren't yet ready to tap keypunchers on the shoulder; neither are the old, reliable punched cards completely ready to be replaced by paper tape or magnetic tape for *input processing.*

With the possible exception of highly specialized (and expensive) "command and control" systems, *request processing* works best with human buffers between the questioner and the data store.

Computer-prepared notification service is a practical method of *announcement and dissemination.* The whole cost of indexing incoming documents shouldn't be charged against these systems, since it has to be done anyway. One of these days a cheap microfiche of a whole document may replace the announcement card as a throw-away item.

User-system relationships will usually be stormy or ennuied. Builders of new systems should get the users involved in the design. Users of any system must be made to realize that they can delegate some of their routine chores; but they have to be given people they can trust. And their complaints must be listened to—and something done about them.

Systems management requires educated managers, as well as educated users. Mechanized systems are neither panaceas nor placeboes; computers create at least as many problems as they solve. Systems evaluation still partakes of black magic.

And over and over again runs the problem of people—people to design a system, to run it, to use it, and to pay for it. We need to get them, to train them, to keep them. And may those systems which are not created for the people, by the people, and of the people perish from the face of the earth. And rightly so.

I think we now know today's state of the art—although much of it is written between the lines of this volume. Tomorrow's art will be better, but it will have to go a long way to beat what we know how to do today.

William T. Knox
White House Office of Science and Technology

Keynote Address [1]

An information service is not an end in itself, but a device by which users can get information to help solve their problems. If an information service is to be successful, it must demonstrate that it can actually give the user the type and degree of service he wants, when he wants it, and at a price he can pay. This is a challenge, because an information service must compete with its individual users' personal sources of information. The user, it must be remembered, is an adult who has spent his life acquiring the information he needs. He possesses a certain pattern of information acquisition and handling that was formed in early childhood, reinforced by formal schooling, and deeply engrained by practical experience. As an adult long accustomed to using books, the user will continue to use books, reports, and journals until he finds it easier and faster to get information through other means, such as asking a fellow employee or relying on a demonstrably effective information service.

I believe that if today's information services are to serve the user as he wishes to be served, they must be dedicated to the active exploitation of recorded knowledge. They must pay relatively less attention to the acquistion and storage of information and relatively more to those parts of the service which will promote the active use of information. Only in this way will the user be convinced enough to change his sources of information. But convincing him is a long, slow process that requires constant, outgoing searching for opportunities in which an information service can be helpful.

It is obvious that the best information service is not necessarily the one that can produce the most or the fastest information. In-

[1] Portions of Mr. Knox's address were extemporaneous. This presentation is therefore an excerpted version of his keynote speech. —Ed.

stead the best information service is one which produces exact information at the time it is needed.

We know that our personal communications channels have load limits, and we automatically and intuitively control the load on these channels so that we usually receive only that information which we can profitably handle to serve our individual goals. We control the number and type of journals, magazines, and newspapers that we get at home and at the office; we adjust our reading speed; we control the number and type of meetings that we attend; and we even control the number of our personal contacts.

In other words, we operate with full recognition that there are limits to the amount of information we can use and that these limits are very small in relation to the total amount of information which is available. Since knowledge or information is of no use unless it is put into the active stream of daily existence, it is easy to see that the human being is the real limiting factor in our information system. Today, we operate in a state of constant overload as regards information input, and our problem is, therefore, to select precisely that quantity and quality of information most effective in enabling us to fulfill our goals and responsibilities.

Planners of information services should keep in mind, therefore, that a newly developed information system may be rejected if it adds measurably to the information overload already faced by the individual, no matter how efficient, glamorous, or fast the new system may be. System designers must also be prepared to experience rebuffs from these people to whom information poses a problem. An old proverb says that "ignorance is bliss," and ignorance definitely is easier to achieve than knowledge. Moreover, the problems are multiplied in the field of scientific and engineering data, where the professional man has been trained in an atmosphere that glorifies the experimentalist and downgrades—sometimes implicitly, sometimes explicitly—the worth of previously recorded information.

Many of the problems in the field of scientific and technical information services have been caused by an undue reluctance on the part of those supplying the services to respond to changed market demands. What is needed is a total marketing-oriented

approach to information services—an approach sensitive to all the marketing variables.

No service has a perpetual hold on the public's fancy—unless it it be medical service, which caters as no other service can to the individual's deep-rooted instinct for physical self-preservation. All other services must continually try to win the public's interest and support. The public in turn assigns to each service a priority, a ranking in relation to other services. When it is remembered that over forty percent of U. S. wage earners are employed in service occupations, it is easy to understand the necessity for any service to enlist the public interest in its support.

How is this done? By total marketing—marketing in the broadest sense. This does not mean selling. In the words of Theodore Leavitt, "Selling focuses on the needs of the seller; marketing on the needs of the buyer. Selling is preoccupied with the seller's need to convert his product into cash; marketing with the idea of satisfying the needs of the customer by means of the product *and* the whole cluster of things associated with creating, delivering, and finally consuming it."

It has been stated that the biggest problem in most information services is in developing the necessary level of financial support. An effective total marketing of these services will, I am confident generate this support.

Let us then look at information service as a business with service as its product—not abstracts, not indexes, not books, but *service*. The Bell System presents an example for us of how a total marketing approach can succeed. It is one business that has truly contributed to the satisfying of its customer's needs and at the same time turned a tidy profit. Note the motto of the Bell System—"Service is our most important product."

The total marketing approach of the Bell System includes advertising, to remind one that others would like to hear from him; thus it crystallizes a latent consumer need. It includes extension phones, to cater to the prevailing custom of taking as few steps as possible. It includes colored phones, to take advantage of the interior decoration emphasis. It also includes phones stationed along streets and highways in recognition of the increasing numbers of people in transit. In summary, the Bell system recognizes the various moti-

vations behind one's use of the telephone and changes its services as the consumer changes his standards of value. The success of the Bell System, with its total *marketing* orientation, may be contrasted with the near-failure quite a few years ago of a large automobile manufacturer guided by a total *production* orientation, which was expressed by a famous saying that customers could have a car in any color, as long as it was black.

Let us then concentrate on the marketing side of information services, too long ignored—not on the production side. The major attention and financial support the information services profession has given to hardware and information processing techniques indicates an over-emphasis on production variables. And by continuing in this way, the information services profession is risking losing more of its market.

Marketing information services in the way it should be done will probably not be easy. It will require new attitudes, new patterns of thoughts, new approaches—probably new people. The record speaks for itself. With only a few exceptions, those in the information services business at the close of World War II did not recognize the dramatic changes that were then stirring in the market. The sweeping national interest in research as an instrument of national security and of economic growth, together with the resultant changes in the market for information services, simply went unnoticed.

What were these changes? There were quantitative changes: the volume of information and the number of R&D clients increased markedly. There were qualitative changes: information was packaged in new forms; clients wanted faster answers, and answers covering a larger number of disciplines; fewer individuals patronized the central wholesaler of information services, more were served directly by an organization's library or information division; the Federal Government became the major supporter of the nation's research and development program. Finally, there were dramatic changes in the technology applicable to information service operations.

It is instructive, if saddening, to review the reaction of the existing information services community to these almost revolutionary changes in their market. The predominant reaction was defensive

in nature and excluded from serious progressive consideration those disturbing features of the new scene which could not be fitted into the existing framework.

As is always the case, the users' needs triumphed. And since their needs were not met by the existing information services community, entrepreneurs more sensitive to those needs found a real opportunity to provide a service, and with profit to themselves.

So much for what is past. What does the future hold in store for all of us in the information service business? Turning again to the total marketing approach, may I suggest that the greatest lack, and therefore the greatest opportunity, is in the broad area of the value of information services.

A sound assessment must be made of the values placed on information services by the user. Latent values, unrecognized by the user, must be made obvious. Hidden barriers to information use must be brought into the open and demolished. Finally, the information services must be organized in direct response to the user's complete spectrum of value judgments.

We have a professional responsibility to make clear to every user the consequences of his combined value judgments on information services. If the system he truly values will be more costly, it is the profession's responsibility to create a greater awareness of the value of information services in order that he will be willing to bear the increased cost. This will call for a massive educational effort. There is, without doubt, a large and growing market for information services. But it is, in my opinion, neither as large nor as fast-growing as might be expected.

There has been much talk and some activity about determining the "real" information needs of the researcher, the engineer, and all the other categories of users. Concern about this problem is all to the good, but I wonder whether the current focus on new and improved information products isn't missing the really essential consumer need. Not that the research and development on new forms of indexes, new forms of citation lists, and new groupings of subject matter is wasted—I feel that it must go ahead. However, that effort is sidestepping a big problem, a truly difficult problem—one of the hidden barriers I mentioned earlier. I refer to the problem of creating in the mind of the consumer the idea that use of

information services ranks among the most desirable, the most valuable functions of the professional man.

Frankly, there are few scientists and engineers who share this view today. The "image" of the user of information services, as well as the "image" of those providing information services, is not one calculated to attract others. This is true although the average professional man values highly a good book or journal article. In my opinion, he doesn't normally extend his value judgment about a specific information service, such as a good book, to the broad spectrum of information services. The training most scientists and engineers receive emphasizes experimentation and sometimes downgrades the value of using literature. The scholar of old, who was the principal user of literature and who therefore set the guidelines for traditional libraries, has no appeal today—he's not lean, vigorous, masculine, sexy. There's a new potential clientele for libraries and other information services which has at present only a nodding acquaintance with books and scholarship. It is this clientele's needs, as recognized and cultivated by the information service professionals, which will govern the shape of future information services.

I am not saying that our problem will be solved by a blitzkrieg public relations campaign. There must be a solid foundation of true value in our information services. But this can be established, I am confident, by continued product improvement together with making the changes that may be necessary in the structure and organization and techniques of the conventional information services business.

A final word or two of a more personal nature. I have temporarily left the cloistered halls of industrial research for the hurly-burly of Federal Government service. There are a number of things which Dr. Hornig[2] hopes to have accomplished during the next year or so. None of them is more important than the establishment of a new conceptual framework for the national information services network which is needed to undergird further progress in

---

[2] Dr. Donald Hornig is the President's Special Assistant for Science and Technology and is head of the Office of Science and Technology of the Executive Office of the President. —Ed.

science and technology and related areas of knowledge. Within this conceptual framework, guidelines can be drawn to help the long-range development of a more effective and efficient information system—federal agencies, library associations, professional societies, industrial research groups, information services groups, publishers, and others.

It is my great fortune to be working with these groups at this time. There is a spirit alive and working everywhere to get on with the job—a spirit of active, dedicated cooperation. With this spirit so dominant, we are bound to take giant strides toward our goal of supplying people with information they need when they need it, in the form they can use, and at a price they can pay.

PRESENTATIONS

Raymond P. Barrett
System Development Corporation

CIRC—Centralized Information Reference and Control

CIRC, the Centralized Information Reference and Control Concept, is a large-scale, broad-spectrum, semiautomatic document-handling system designed and implemented by System Development Corporation under a contract with the United States Air Force. The title of the system is meaningful in that CIRC is indeed a central reference concept—not a central storage concept—and does permit central control as that word applies to large-volume dissemination of materials having many different classifications and releasability restrictions. CIRC is truly a large-scale system in that it must process about 8,000 to 10,000 input reports per month, provide a retrieval capability for the total system holdings, and carry out weekly dissemination to a large number of user groups scattered throughout the United States. Further, CIRC is a broad-spectrum system which involves indexing, classification, abstracting, translation, conversion and storage, retrieval, and dissemination. Several of these functions are not automatic operations and others are being studied for possible conversion to machine-processing procedures in the near future.

CIRC is newly operational, actually only partially operational, and therefore is still not unchangeably rigid with respect to operating procedures. We hope that by the time this workshop is concluded we shall have profited from the composite operational experiences of this group and that we can enhance the CIRC concept by increasing its capabilities, reducing its costs, and further refining its responsiveness to the needs of the engineers and scientists who call upon it for support.

It is not reasonable, or even desirable, to attempt a detailed assessment of CIRC design specifications in this short and generalized presentation. I do believe, however, that is will be useful to

summarize that information in terms of some general comments on system objectives.

System Objectives

It was determined that CIRC should provide a single-point, total data-base interrogation capability. The pre-existing data base was comprised of many different, mostly manual, subsystems, situated in a variety of physically separate locations, some quite remote from the main body of users. Most of the pre-existing files were organized on different principles—some by type of material, some by source of material, some by subject matter, others by country, and so forth. The user agency desired, and required, single-point access to the entire data base holdings. In support of that requirement it was necessary to design and implement a single master index system to replace the various dissimilar systems used in support of the multifile situation. Further, it was required that the new concept be built around a controlled indexing vocabulary which could be used to minimize language ambiguity with respect to the master index and the storage, retrieval, and dissemination functions of CIRC. It was further stipulated that search response-time be changed from the highly variable and often lengthy situation, which characterized pre-existing efforts, to a matter of dependable, total-file search on a 24-hour basis—over-night service. Provision for user control was also required so that the user could control the search and screen capabilities of the system in such a way as to prevent his being inundated with an undifferentiated, i.e., unranked, mountain of reports more-or-less responsive to his requests. The most efficient conceivable document retrieval system could fail to satisfy this user because of these reasons and because of the very large data base holdings and data base acquisition rate. A mode of user-controlled "levels of access," that is, user-controlled search and screen capability, had to be provided.

With respect to dissemination, it was determined that the geographically widespread network of users, which was characterized by diversity of mission, of staffing, and of subject interest, could not be served by a dissemination system which provided only a single kind of output at a single level of access (for example, subject detail). A high-volume, user-responsive, multi-level dissemina-

tion system operating on a weekly basis was required. It had to provide a spectrum of service, from dissemination of specialized accessions listings, to profile-controlled notification, proxy dissemination, and, finally, parent document dissemination. Of highest importance, the system had to have a method for keeping user requirements very specific and constantly updated.

Finally, it was important to build the new system capability in such a way as to establish communications compatibility between CIRC and certain information systems serving the military and other parts of the Federal Government. By and large, this meant careful coordination of the vocabulary effort with the Defense Documentation Center (DDC), among others, as well as attention to concepts of data base structure being evolved by the Department of Defense.

Design Constraints.

With respect to design constraints, I would like to limit my remarks to those which may be out of the ordinary with respect to many other system efforts in semiautomatic document processing.

It was decided at the very outset of the effort that we should build a system that provided the user the most responsive service possible with respect to the system's objectives. But it was also required that the spectrum of machine functions be based upon present state-of-the-art capabilities. Every subsystem, every technique, had to be one that had been tested, at least to the level of technical feasibility, somewhere within the information processing community. The success of the resulting design was to be in no way dependent upon expected breakthroughs or upon the design of new, specialized, sophisticated equipment. This is not to say we were to inflexibly ignore new and promising techniques; we were simply not to depend on the chance that they might succeed. CIRC, then, was not to be, and is not, a research effort: it is an applications program.

With respect to equipment, CIRC was constrained in a practical sense to an already available configuration. This meant a Flexowriter for input, IBM 7094 and 1401/1403 computer facilities, and a complex of 16-mm Lodestar/Recordak film storage, retrieval and printing devices. I shall describe later how we made use of these equipments.

Implementation presented some interesting special problems. First, it was necessary to bring the system into being by a series of increments in capabilities. One subsystem at a time was phased from design and development into operation. Then, it was necessary to run each subsystem in parallel with the pre-existing procedure for a period of time in order to insure that the new procedure was fully accomplishing its purpose. The user personnel decided when to cut off the old procedure and depend upon the new one. There is, in my opinion, no better mechanism than this for obtaining the confidence and support of a user.

CIRC implementation is input oriented. Each subsystem is that set of programs and procedures required to move a given kind of input material through each of the processing steps or elements which comprise the total system operation. I personally prefer this approach; but I also recognize its inherent problems, perhaps the greatest of which revolves around the necessity to accept continued system program modification until all input types have been incorporated and to anticipate a program clean-up and optimization phase at or near the end of the total system implementation effort.

A FUNCTIONAL OVERVIEW

It will be well to establish a general frame of reference concerning CIRC in order that we may examine its constituent functions in some greater detail. (See figure 1.) All externally and internally

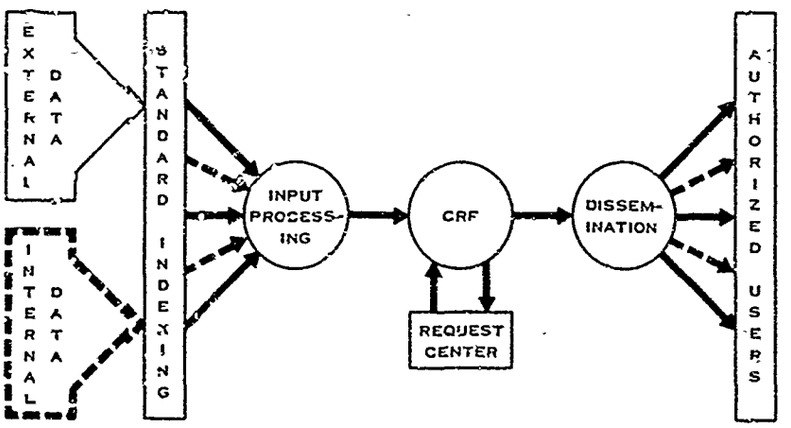


Figure 1

generated reports judged worthy of retention are subjected to standard indexing and classification–so-called descriptor-word indexing. ard indexing and classification-so-called descriptor-word indexing. The resulting materials are converted to machine form and stored in a Central Reference Facility (CRF) in an inverted coordinate index. The heart of the CRF is a 7094 computer supporting the Request Center. A semiautomatic dissemination system distributes several different levels of information to both internal and external user groups all over the nation. These three major areas, input processing, the CRF, and dissemination, constitute our frame of reference, and each merits some additional individual consideration.

Let's look at the nature of CIRC's data store (fig. 2). Every item that enters the system is caused to exist on three levels: a header data level, a proxy level, and the parent document level. The header data level is comprised of the standard bibliographic reference information: the title, author, source, date, classification, and so forth. Along with this there is a group of "topic tags." These

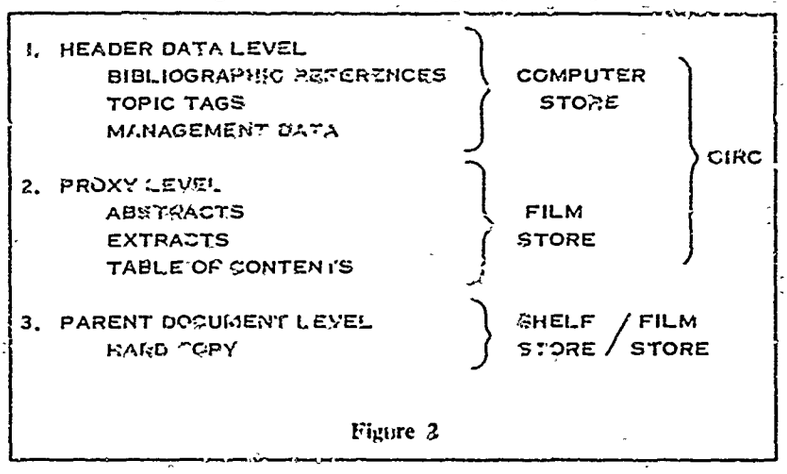


Figure 2

are one-, two-, or three-word expressions which characterize the nature of the document's contents. They are called "clue words" or "key words" or "uniterms" by other organizations. We chose to call them "topic tags" so that we could define them clearly and avoid confusion with concepts which apply to other systems but not to CIRC. Topic tags, then, are used in a form of descriptor indexing.

We also have some items of management information in the header data. If you are going to run an information system on a computer, one of the bonuses you can get is a management system which tells you, for example, how many of each kind of report have been received this month versus last month, how many are put out, where the bottlenecks are, and so forth. This is a business system such as any large organization, military or civil, runs to keep track of its operations. The bibliographic references, the topic tags, and the business data which comprise our depth index are computer manipulated.

Next, we have the "proxy." Now, a proxy isn't a thing—it is a level of information. It tells more about the contents of a document than does the header data; but it is easier to screen, easier to handle, than the parent document. For us, a proxy may be an abstract, or it may be an extract, or, in the case of a large reference book, it may be the table of contents. The proxy level of information in the CIRC system is stored on film. The parent documents themselves remain in hard copy in shelf storage, although they also are gradually being reduced to a microfile.

As we move on now to the functions of the Central Reference Facility, I will show you why we need these levels of information representing each input document.

CENTRAL REFERENCE FUNCTIONS

CIRC operates in a number of retrieval modes. First, the system has to be able to retrieve in terms of a subject-specified search. This is a matter of a user approaching the system and describing the subject area of concern. He might talk about metals and stress and cryogenics, and so on. Having done so, however, he generally moves on to elaborate his subject-specified query in terms of qualifiers. He may say, "I don't want anything older than 1958" or "I don't want documents concerning American research," or he may say, "Documents must all be unclassified."

CIRC also must be able to search on title, author, source, date, and collations thereof. It must be able to retrieve the proxies, the middle level of information, and, since it is not a storage center, it must be able to locate parent documents when they are required. Lastly, it must provide a capability for retrieving management data statistics.

I think it would be of use now to look at the example of a subject-specified-and-qualified request in figure 3. The input topic tags are

```
THE FOLLOWING REQUESTED TERMS FROM REQUEST - - - RQSTI

   METEOROLOGIC ROCKET          END-BURNING ROCKET          SMALL ROCKET
   SOUNDING ROCKET              ATMOSPHERIC SOUNDING        WEATHER FORECASTING
   SOLID ROCKET                 ATMOSPHERIC SAMPLING

HAS YIELDED 0100 DOCUMENTS FROM THE CIRC DATA STORE.

THE FOLLOWING 0100 DOCUMENTS MEET ALL SPECIFIED QUALIFICATIONS. . . .
                                                    QUALIFIERS
                                                       DATE, . . . . . .NONE
                                                       SUBJ CODE . . . . .NONE
                                                       TYPE DOC, . . . . . .NONE
                                                       CLASS, . . . . . .NONE
                                                       RELEAS, . . . . . .NONE
                                                       COUNTRY . . . . . .NONE
                                                       MUST TERMS . . . . .NO
                                                       NUM HITS . . . . . .SINGLE AND MULTI-HITS
                                                       NUM DOC . . . . . .300 MAX OUTPUT
                                                       TYPE OUTPUT. . . . .A

05 HITS
   ACCESSION NUMBER . AP4000443     DATE SELECTED. . . . . .03/01/65     CLASSIFICATION . .UNCLASSIFIED NUM 001
   COUNTRY. . . . . .USA            SUBJECT CODE. . . . . .ES            DATE PUBLISHED  .1963
   TYPE DOCUMENT. . .JOURNAL        CAN NUMBER. . . . . .AA04/0376
   NUMBER PAGES. . . .006           REEL & FRAME NUMBER .65/20/57
   RELEASABILITY. . . .NO RESTRICTIONS
   TITLE. . . . . .SMALL ROCKETS

   AUTHOR . . . . . .DANIEL, O. H.
   PERIODICAL . . . . . .INTNTL SCIENCE AND TECHNOLOGY, V. 40, APRIL 1965
   TOPIC TAGS . . . . . .METEOROLOGY, METEOROLOGIC ROCKET, UPPER ATMOSPHERE, SOUNDING ROCKETS *
                         ATMOSPHERIC SAMPLING, * ARCAS, SOLID ROCKET, ATMOSPHERIC SAMPLING
```

Figure 3

either proposed by the requester or selected from a written request provided by him. With the available qualifiers it is possible to qualify the request by date, by subject code, by type of document, by classification, by releasability, by the country concerned. The "must terms" qualifier is simply a mechanism to allow a limited Boolean-type retrieval capability.

"Num hits" applies a criteria governing the amount of output in response to this request. In effect, it is possible to specify the number of matches between input terms and indexing terms that must be present before a document candidate is considered a "hit", that is, respo. e to a request.

"Num doc" allows the request to be qualified so that it results in only the best ten, fifty, or hundred documents; or, if you want, all the responsive documents in the data store. This is of some importance in that the output from the system is, in an elementary way, rank-ordered with respect to relevance of the request.

"Type output" is a three-level parameter which allows the requester to govern the amount of information he wants printed out concerning each retrieved report.

One of the first things the system does is tell how many documents in the store are responsive to the request. You will note that there were 100 documents responsive to this straightforward request. We think it is important to notify the user how many documents are available in order to allow him to retrieve the material in rank-ordered blocks rather than in some sort of database dump. Although our data base is now quite large and increasing rapidly, the time will come when the number of responsive documents will total hundreds or even thousands.

Let's look now at the type of material retrieved. It's our old friend, the header-data level of information—the accessions number, the date selected, the classification, the date published, the country, the subject code, the type of document, the number of pages, releasability, the title, the author, the periodical, and the topic tags. Note the number "05." This tells us that five of the input terms matched five of the indexing terms in his document. If you could see the rest of the printout you would see this decreasing from five to four, to three, to two, and so on. The output is rank-ordered by counting up the number of matches between

input terms and indexing terms. A user, looking at this kind of demand bibliography, can screen it rather rapidly. He can select those articles which are unknown to him or which look interesting to him, and he can see their abstracts by punching their reel and frame numbers on the keyboard of a Lodestar device. He can then screen the abstracts and ultimately ask for a very small number of hard-copy parent reports.

Let's look at figure 4 for a better picture of how this works. We'll start off here with the nature of a request—the substantive

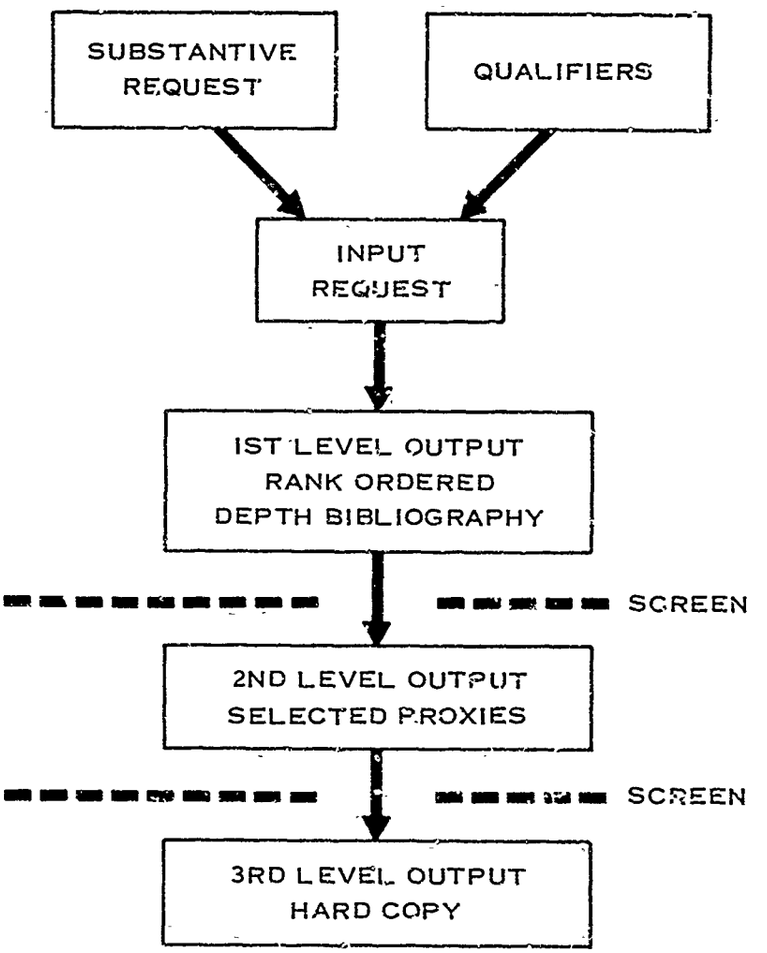


Figure 4

request; i.e., the topic tags that characterize the nature of the request and the qualifiers. Remember, a requester can qualify by date, by country, by classification—quite a number of things. He can decide whether he wants to use simple "or . . . or . . . or" logic or a limited Boolean combination. This information is input to the system, and the first-level output is a rank-ordered bibliography. He screens the bibliography and then uses the reel and frame number to look at the second level of output, the selected proxies, or abstracts. He screens the proxies and then asks for a very small number of the third-level output, the hard copy.

The procedure is based on the fact that our user group has a tremendously large data base. The best conceivable and most efficient conceivable document retrieval system operating on a single level of output would stack documents in some way responsive to his request about neck-high around the requester's desk. The user unfortunately would never have time to read through all of that material. Therefore, such a document retrieval system would not be a useful tool for helping him with his job. CIRC, on the other hand, provides information in such a way as to help the user screen through the chaff and arrive at a small number of potentially very useful hard-copy documents. Now, this output mode is not forced on a user. He has the option of asking for a demand bibliography and stopping there. In another case, a user may want to go directly to the proxy level, to the abstract, and he can do that. In yet another case, he might insist on going directly to the hard copy. Unless we knew that there was only a very small number of responsive documents in the file, we might attempt to dissuade him from this because of the bulk problem, but the option belongs to the user, and this screening process is a tool at his disposal *if* he desires to use it.

Now for a look at one of the other major products of the Central Reference Facility, the selective, collated accessions listing. When we started developing CIRC, we found that the user group had been receiving monthly accessions lists on each of the several different kinds of input materials. The users complained that there were too many accessions lists and that they were too large. They felt, however, that it would be a mistake to discontinue accessions lists altogether. They were useful when one could find enough

time to study them. We were asked what CIRC might do to help this situation. Since all of the material is flowing through a machine system, it seemed easy enough to say, "Well, we can produce one consolidated monthly accessions list instead of having separate accessions lists with different kinds of data." The difficulty here of course concerns the size of the resulting book—and it is a book, several inches thick. Some way had to be developed so that the user would not have to wade through all of this material in search of the few items of concern to him.

Here is the solution we implemented. If you opened an accessions list, one of the first things you would see is a list of thirty-three subject areas (fig. 5). Now, these areas are selected out of

AA	ASTRONOMY AND ASTROPHYSICS
AC	AIRCRAFT AND AERONAUTICS
AS	AEROSPACE STRUCTURES
CB	CHEMICAL, BIOLOGICAL AND NUCLEAR WARFARE
DC	DETECTION AND COUNTERMEASURES
DP	DATA PROCESSING INFORMATION SCIENCES
EC	ELECTRONICS, COMMUNICATION AND COMMUNICATION EQUIPMENT
EE	ELECTRICAL ENGINEERING AND ELECTRIC EQUIPMENT
EM	ELECTRICITY AND MAGNETISM
ES	EARTH SCIENCES AND PHOTOGRAPHY
FP	FUELS, PROPELLANTS AND LUBRICANTS
GC	GENERAL CHEMISTRY AND CHEMICAL ENGINEERING
GM	GUIDED MISSILES
GO	GOVERNMENT, COMMERCE, CULTURE, ETC.
GP	GENERAL PHYSICS AND ACOUSTICS
IC	INORGANIC CHEMICALS
IE	INDUSTRIAL ENGINEERING
LS	LIFE SCIENCES
MA	MATHEMATICS
ME	MECHANICS OF RIGID BODIES, LIQUIDS AND GASES
MM	METALS AND METALLURGY
MS	MILITARY SCIENCES
MT	MATERIALS (NON METALLIC)
NG	NAVIGATION AND GUIDANCE
NP	NUCLEAR, ATOMIC AND MOLECULAR PHYSICS
OC	ORGANIC CHEMICALS
OP	OPTICS
PH	PSYCHOLOGY AND HUMAN ENGINEERING
PR	PROPULSION
SS	SOLID STATE PHYSICS
SV	SPACE VEHICLES AND ASTRONAUTICS
TD	THERMODYNAMICS
WA	WEAPONS, AMMUNITION AND EXPLOSIVES

Figure 5

the experiences of the operating center and have no validity outside of the area of that particular user group. A requester now can select one, two, three, five, ten, or whatever number of these general subject areas he wants, and we will send to him each month an accessions list covering only his areas of interest. You can see how this cuts down the size of the book a requester has to wade through. This is a big help but is not a complete answer. If you turned the page in the accessions list, you would see a table of contents like the one in figure 6. This table of contents shows that

TABLE OF CONTENTS

Figure 6

six subject areas have been selected; and you will see that under each subject area we have collated the data by input type. Now, a user can get information only in the subject areas of concern to him, but he can very rapidly select out and read only the kinds of input that he thinks are of value to the work he is doing. If he doesn't want journal articles, he can skip them. If he is interested only in field studies, he can look only at field studies in the five areas of interest to him. The level of information printed out in an accessions list is based completely on the header data level—the author, title, source, descriptors, accessions number, references, etc. We have, in effect, produced a selective and collated accessions list in response to the user's needs.

DISSEMINATION

Let us turn now to the last major functional area of CIRC, dissemination. All of our dissemination is semiautomatic (that is to say, it is computer-supported) and depends upon the existence of very specific and very detailed user profiles. A user profile is a list of topic tags, or descriptors, which describe the scope of a user group's interest. Note that I said *group*. All of our dissemination is based upon unit profiles. We discovered early on that individual profiles had a very high degree of duplication and that it appeared more economical to talk in terms of a profile serving a unit rather than an individual. Such a unit might have two, three, five, or even ten people in it working on closely associated subject areas. As a matter of experience, we discovered that a unit profile averaged out at about 1,000 of our precoordinated topic tags. The system has been running for about a year now and is currently serving 70-odd user groups scattered across the United States.

There are several different kinds of semiautomatic dissemination. First off, profile-controlled notification. Some of our customers desire to be kept aware of everything flowing through channels that is responsive to their area of interest, but they do not wish to have the actual materials sent to them automatically. These people receive a notification once a week that consists of the header-data level of information on each article that came into CIRC which fits their use pattern and which is responsive to their profile. Other users inform us that they prefer proxy dissemination,

in which case they receive the abstract or extract—the proxy level of information—on each document responsive to their profile. Some customers prefer to receive both, because the notification is very easy to screen and the proxies are useful as references. We found also that it was necessary to provide each user group with a control for volume and specificity. These user groups vary from just a few people to dozens of people; therefore, the capability to assimilate raw data varies accordingly. There is a clip level in the program—a little formula—by which we can vary the machine criteria which determine what shall be sent to a specific customer, and a different setting can be applied for each user group. With respect to specificity, there is a feedback program, primarily an automatic one, which makes small demands on the users on a routine basis. Finally, we can consider the dissemination of the selective, collated accessions list to be a type of dissemination which serves a different level of our information users.

It might be useful now to examine the kinds of information that are sent out to our user groups in the weekly dissemination of notifications. Figure 7 shows an excerpt from a notification. You

ACCESSION NUMBER– AP4000443 COUNTRY – USA ( ) NOTICE IN INTEREST AREA
DATE SELECTED – 03/01/65 SUBJECT CODES – P,S ( ) NOT IN INTEREST AREA

TITLE – SMALL ROCKETS

TOPIC TAGS – METEOROLOGY, METEOROLOGIC ROCKETS, UPPER ATMOSPHERE, SOUNDING ROCKETS,* ATMOSPHERIC SAMPLING,* ARCAS SOUNDING ROCKET, ATMOSPHERIC SAMPLING

AUTHOR – DANIEL, G. H.

PERIODICAL – INTERNATIONAL SCIENCE AND TECHNOLOGY V. 40, APRIL 1965

Figure 7

will notice that it is the header-data level of information—accession number, the date, the country, the subject codes, title, the group of topic tags, the author, and the periodical. Note that after some of the topic tags there is an asterisk. This informs the user that the article was called to his attention because these topic tags occurred in his profile—"sounding rockets" and "atmospheric sampling." If the user so desires, he can cause these words to be deleted from his profile, or he can cause other topic tags to be added to his profile. In other words, the user can respond with feedback on individual articles which modifies his profile.

A quarterly, automatic review of his profile is accomplished in a different way. Once every three months we cause the machine to print out the user's profile, and next to each word in the profile we print a number showing how many times that word was useful in getting him an article. As you might guess, we pay particular attention to those words which were not working during the past quarter and to those words which were overworking. This means that once every quarter a man from CIRC visits each of our dissemination system customers and reviews his profile and its performance over the past three months.

Originally, notifications were prepared on standard printout paper, but starting in June (1965) we are changing to card stock—one card per notification. This provides the recipient with notifications which can be filed in a standard, manual file box under the indicated subject headings, thereby giving the user some capability for retrospective searching of materials selected for him by the semiautomatic dissemination program.

SYSTEM TRAINING AND TURNOVER

One of the prime keys to the successful design, development, and implementation of any system has to do with indoctrinating and training the ultimate system user. Indoctrination and training of different kinds should begin as early as possible in the implementation cycle. We have found it necessary to institute classroom programmer training, to train government employees as computer programmers, followed by on-the-job programmer training. This means that new government programmers work across the desk from experienced SDC system programmers in design, development, and checkout of the actual programs used in the CIRC information system. As a point of fact, some trainees have actually written some of the less-complicated programs which are currently being used in the CIRC operations.

One other kind of training worth noting that has proven quite effective has to do with on-the-job training with respect to the total information system. Running an information system involves many other things besides computer programs and a computer, since many of the steps are manual operations. On-the-job training in indexing and classification, for example, involves professional hu-

man activities which provide the logic links between computer operations, system operations and effective communications with the user groups.

CURRENT STATUS

CIRC is not completely operational. All of the basic programs are running and all of the basic capabilities are in existence (see fig. 8), but CIRC is not yet accessing the full load of input materials

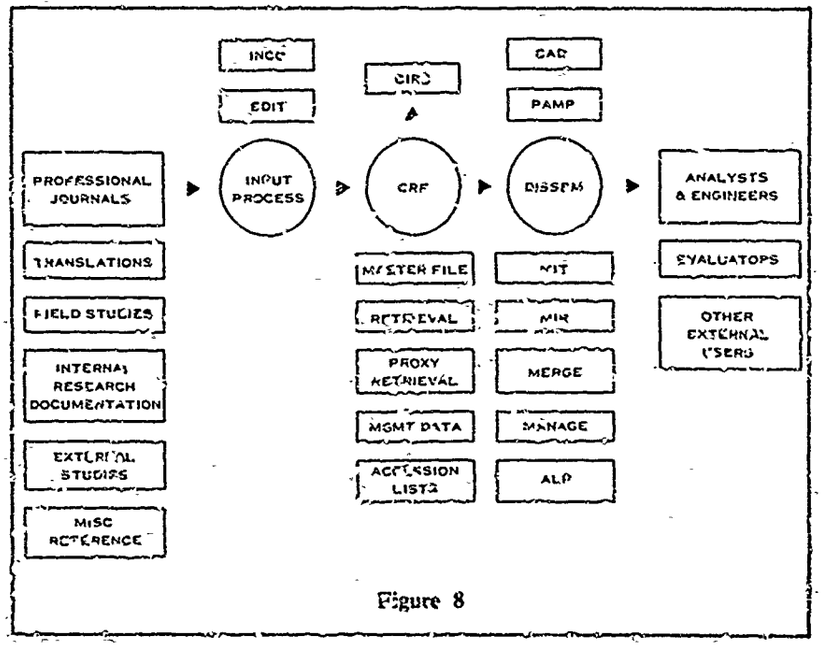


Figure 8

that it must eventually process. Of six different types of input materials, only the first four presently are flowing into the system. Of a total of approximately 8,000 to 10,000 input items per month, CIRC presently is processing only about 4,000. We still have a way to go to come up to full speed.

It might be of value now to run through some of the basic programs. INCO is a program that looks at the punched paper tape input after it is converted to magnetic tape, detects certain mechanical errors, and outputs them for correction. EDIT acts as a switching center; it sends part of the data on the input tapes to the master file, part to various retrieval files, part to the management data files and part to the accessions list files. MIT builds and updates

the master file. MIR is a retrieval program. MERGE puts the reel and frame number of the proxies onto the MIR output. MANAGE is a set of little programs that produce management data reports. ALP produces the selective, collated accessions lists CAD accomplishes semiautomatic dissemination. And PAMP produces quarterly profile analyses (that is, profile feedback outputs).

PLANS AND PROSPECTS

Finally now, let us look at some of the things presently under study and consideration as future CIRC capabilities. First, we are interested in coming up to full speed; that is, processing the full 8,000 to 10,000 items across the six different subsystems. Next, we are looking at a variety of presently existing auto-extracting programs to determine whether any one of them or a modification of one of them could be used to produce the proxy level of some of the materials required as CIRC input. I am pretty sure the answer is positive in this case, but we are moving slowly and with great caution. Next, we are beginning to prepare for converting CIRC from a basically tape system to a disc system in the interests of maintaining speed of operations in the face of a dramatically increasing data-base size. We are also looking at some list-processing techniques which may be used to modify the logic of an inverted coordinate index in such a way as to facilitate searching while minimizing search time. We have launched a program to examine the existing historical data base in terms of determining the most efficient way of converting some subset of it to a CIRC system. Last, we are looking at time sharing for remote inquiry capabilities for the CIRC system. It appears feasible in a technical sense and an economical sense for multiple users to query the same computer file from remote teletype consoles.

C. Allen Merritt
International Business
Machines Corporation

An Operating System The IBM Technical Information Retrieval Center

A very brief description of the IBM Technical Information Retrieval Center is that it is a centralized in-house computerized information retrieval and dissemination service designed to satisfy the scientific and technical information needs of IBM's scientists and engineers. This activity, located at the Thomas J. Watson Research Center at Yorktown Heights, New York, was formed in November 1964. This central organization brought together the professional personnel, the system knowledge, and a large data base that could best serve the domestic and foreign scientific and technical community of IBM. At the time the Center was being organized, a decision was made to use the normal-text information retrieval technique, developed in IBM, as the basis for operating the system.

The normal-text searching technique provides fast access to research and engineering information of interest. With the large volume of reports, the frequent changes, and the interdisciplinary nature of our many research and development projects, it is necessary that a thorough search be made of the data base each time a query is answered. Ordinary English sentences provide a natural and convenient medium for expressing factual data. Normal text is understood, used, and accepted by scientists, engineers, and administrators for information and reporting purposes. Searching the textual data eliminates the usual time-consuming and error-introducing classification coding, key wording, or structuring problem involved in many IR systems. Finally, normal-text searching allows the data base to be searched efficiently, rapidly, and economically. Information cannot be considered a luxury; however, avoidance of high cost and delay is essential not only to the Center's activity but also to the user.

In addition to the Center located in Yorktown Heights, we maintain three active satellite operations at San Jose, California, at Endicott, New York, and at our World Trade Laboratory in La Gaude, France. Each of the satellites has a complete set of our data base tapes and the necessary programs for normal as well as emergency searching.

Some of the types of input information that we process include those listed in figure 1. IBM Research and Engineering Project Files are the official reporting medium for all Research and Development activities within the Corporation. IBM Technical Reports are prepared by the engineers and scientists as internal company-confidential reports.

IBM RESEARCH AND ENGINEERING PROJECT FILES
IBM TECHNICAL REPORTS
IBM INVENTION DISCLOSURES
IBM RELEASED PUBLICATIONS
IBM SUGGESTIONS
TECHNICAL PRESS LOG
NON-IBM TECHNICAL REPORTS
SELECTED U. S. PATENTS
ABSTRACT-INDEX BULLETINS
CURRENT INFORMATION SELECTION
CENTRALIZED MICROFILMING
RETROSPECTIVE SEARCHES

Figure 1

Often these are cleared at a later date for outside publication or distribution. IBM Invention Disclosures are submitted as novel ideas to solve specific problems. The most promising of these are selected to be filed as patents. IBM Released Publications are papers or reports that are presented or distributed outside IBM and by IBMers. Non-IBM Reports are selected from many journal and magazine articles, government reports, and various university and college reports and theses. All input to the IBM Technical Information Retrieval Center that is not originated within the Company receives the necessary copyright clearance before it is placed into the system. In IBM, as in many large companies, we have a

suggestion system for recommended changes in procedures or products. New suggestions are compared by computer with previous suggestions to determine whether a previous suggestion covering a particular idea has been submitted. If no match occurs, the suggestion is investigated further to determine if it should be accepted for implementation and an award. We also have on tape a selection of U. S. patent claims to determine the feasibility of assisting in infringement searching. The complexity of searching here is a degree higher than in the usual information retrieval search.

Services of The Center

The services of our Information Retrieval Center include:

Retrospective Searches. A complete search of current and historical files tailored to a requester's needs. At the present time we have over 125,000 documents that we can search in total.

A Current Information Selection Program. A search of the new data continuously being added to the data base to provide our scientists and engineers with current awareness of new information by matching the individual user's profile with the data base. A profile is composed of the technological terms, sentences, or phrases used to describe the individual's professional assignment. A match between the data base and the profile results in the selection of items to be directed to a person.

Abstract-Index Bulletins. An on-the-desk browsing tool for the engineer or scientist containing abstracts of documents, a category index, and a subject index. These are published monthly and contain the monthly acquisitions arranged for ease of handling and review.

Centralized Microprocessing. A Center service to provide each IBM Technical Library in over 26 U. S. and foreign locations with backup copies of reports in the ITIRC System. Hard-copy needs are serviced by the local IBM Librarian. At the present time there are over 24,000 reports on microfilm.

Specialized Librarian Tools. As a natural by-product of the system, special runs and cumulative listings are processed and distributed to the IBM Technical Librarians. Author indexes and source code listings are examples of this kind of capability. These are all on-the-shelf tools for librarians.

Automated Project File. This is a monthly compilation of all the current research and development projects in the Company. It is distributed to a controlled list of managers as an awareness and control tool. The data are updated monthly and contain descriptions of the project, budget figures, manpower and planning information.

Retrieval by Normal-Text Technique

Let us now take a look into the normal-text information retrieval techniques. The user, or requester, forwards to the Center a question or inquiry consisting of pertinent words and phrases that might be found in our data file. This information is directed to the information specialist. Here is the weakest and yet the strongest point in our Center's activities. That is, the information specialist must be able to interpret the inquiry properly so that he matches the needs of the user with the questioning logic and techniques available.

The tools that the IR specialist uses are individual words, phrases, and sentences with contextual logic. In phrasing the inquiry the specialist may use a string of words, that is, a specific linear sequence of words. He may use word combinations. He may use an "or" possibility, to instruct the computer to look for any one of several possibilities of desired words or expressions. He may use an "and" logic combination to instruct the computer to look for all desired words or expressions if an individual were interested in determining all references or examples of a desired subject. He may use a negative determination; that is, he may instruct the computer to ignore specified data or specified words. He may also use an "absolute yes," or a positive answer, so that regardless of the search request or matching criteria the computer would find and print out each case. This is called an override in our system and will print any answer found regardless of other logic or criteria used.

The adjacent word is probably one of our most common tools in preparing a search request. If one were interested in a specific subject and wished to avoid false answers, resulting from the "right words" being in the text but having different meanings due to contextual positioning, the computer would be instructed to register an answer only when the words are adjacent. This technique is also useful if an individual is interested in a specific phrase, sentence, or a quotation, such as the very common "to be or not to be."

The logic I have described to you is used to fulfill various retrospective search requests. However, let me emphasize that this same

IBM TECHNICAL INFORMATION RETRIEVAL CENTER
THOMAS J. WATSON RESEARCH CENTER
YORKTOWN HEIGHTS, NEW YORK

RF GALLECK POK DPD 094156
DEPT 798, BLDG 005

IN 00. 1209. COMPUTER LIST PROCESSING LANGUAGES. NOVEMBER 1964.
US-PK

BARBIERI, R.

IN-00.1209

LIST PROCESSING LANGUAGES ARE DESIGNED TO MANIPULATE SYMBOLIC DATA, WHEREAS PROGRAMMING LANGUAGES SUCH AS FORTRAN ARE ORIENTED TO NUMERIC COMPUTATION. THE DATA FOR LIST PROCESSING LANGUAGES ARE SYMBOLS THAT MAY HAVE OTHER THAN NUMERICAL MEANING, AND THE NEED ARISES FOR A UNIT OF DATA LARGER THAN A SINGLE NUMBER. THIS REPORT DESCRIBES SEVERAL LIST STRUCTURES- SIMPLE, LISP, MSS, THREADED, KNOTTED, AND SYMMETRIC- AND SOME IMPORTANT ASPECTS OF THEIR IMPLEMENTATION. THEN IT DESCRIBES AND COMPARES FOUR LANGUAGES- FLPL, IPL-V, LISP, AND SLIP- FOR WRITING PROGRAMS TO SOLVE PROBLEMS SYMBOLICALLY. IT ALSO GIVES EXAMPLES OF THE USE OF SLIP. LIST PROCESSORS HAVE BEEN USED, AMONG OTHER THINGS, TO PROVE GEOMETRY THEOREMS, SOLVE SYMBOLIC DIFFERENTIATION AND INTEGRATION PROBLEMS IN CALCULUS, TRANSLATE ONE NATURAL LANGUAGE TO ANOTHER, AND SOLVE PROBLEMS IN MANAGEMENT AND BEHAVIORAL SCIENCES. 32 PAGES.
A DISSERTATION IN THE DEPARTMENT OF MATHEMATICS SUBMITTED IN JUNE 1964 TO THE FACULTY OF THE GRADUATE SCHOOL OF ARTS AND SCIENCES AT NEW YORK UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

21-PROGRAMMING PROGRAMMING LISTS
PROCESSORS LANGUAGE LISP
IPL RETRIEVAL SEARCHING
STORAGE PROGRAMS FLPL
SLIP PROCESSING
6JA000071-MF001

Figure 2

logic is used in our Current Information Selection Program in which we now have over 1,000 individual profiles on tape.

Let us go to the computer now and assume that we have prepared and placed into it a search request, or that our Current Information Selection Program has been run with a thousand users against 500 or 600 input documents. What type of information do we receive back from the computer? What does the user receive? Specifically, we receive on each document hit the bibliographic information of the document; that is, the source of the document, the title, the author, the abstract, micro-film control number, and subject terms assigned to the document. These subject terms are used for the bulletin subject index (see fig. 2). With this information, a user can then go to his local library and see on microfilm the complete document, or he may request a hard copy of the document directly from his library.

Measurement and/or evaluation of the Center's activities are of primary importance. Whether the information is forwarded to the user as a part of our Current Information Selection Program or in response to a retrieval search, there is a measurement technique that we use for each document called a response card (fig. 3). The response card for the Current Information Selection Program is returned to us after the user punches one of five Port-a-Punches on the card to indicate the relevancy of the document information we have provided to him and whether he desires a copy of the

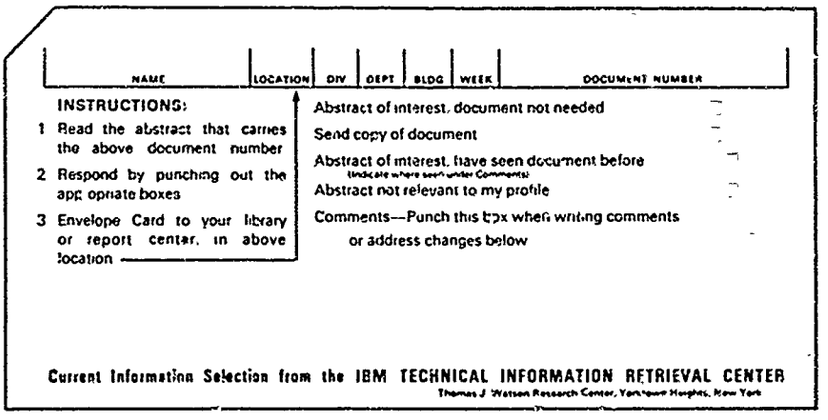


Figure 3

document. When a search response is sent to the user, he indicates to us whether the response was specific to his needs or whether he desires more information.

SUBJECT INDEX

PERFORMANCE
TR-2a, 029, DATA COMMUNICATION SYSTEM DESIGN AND EVALUATION, JANUARY 1965.
65A000443-MF006 PAGE 293

PERSONNEL
64-965/0001H, SATURN IB/V INSTRUMENT UNIT HUMAN ENGINEERING PROGRAM PLAN, DECEMBER 1964.
65A000445-MF006 PAGE 294

PHONONS
NC-416, SOLUTION OF THE GAP EQUATION FOR /PB, /HG, AND /AL, DECEMBER 1964.
65A000198-MF003 PAGE 167

PHOSPHORUS, FILMS
TR 00.1234, THE RELATIONSHIP BETWEEN COERCIVITY AND THE STRUCTURE AND COMPOSITION OF ELECTROLESS /CO-P/ FILMS, DECEMBER 1964.
65A000448-MF006 PAGE 295

PHOTOCOMPOSITION
TR 21.139, PHOTOCOMPOSITION AND AUTOMATED TYPE SETTING, DECEMBER 1965.
65A000447-MF006 PAGE 295

PHOTOCONDUCTORS
TR 07.062, A SURVEY OF COMMERCIAL SEMICONDUCTOR PHOTOSENSITIVE DEVICES, DECEMBER 1964.
65A000195-MF003 PAGE 166

PHOTODIODES
65A000188, INTEGRATED ELECTRONIC GATING SYSTEM FOR MULTIPLEXING APPLICATIONS, DECEMBER 1964.
65A000188-MF003 PAGE 161

TR 22.146, LIGHT COUPLED DIODE, READ ONLY MEMORY, DECEMBER 1964.
65A000449-MF006 PAGE 296

PHOTODIODES, SILICON
NC-454, A SURVEY OF RESPONSIVITY AND SPEED OF SILICON DETECTORS TO /GAAS/ DIODE RADIATION, NOVEMBER 1964.
65A000216-MF003 PAGE 177

PHOTOGRAPHY
EN 64-046, THE PROPERTIES OF ELECTROFAX PAPER, AUGUST 1964.
65A000179-MF003 PAGE 156

Figure 4

As indicated earlier, we provide three monthly bulletins of our current input: IBM Documents, non-IBM Documents, and Invention Disclosures. Each bulletin has three main sections:

1. **The category listing.** This offers a quick way for the reader to scan the contents of the bulletin selectively. This section lists all documents under one or more broad subject headings with titles and accession numbers. There are 23 categories.
2. **The abstract section.** Reports and other documents originated within IBM or selected from outside IBM are identified by an assigned accession number as well as the original source code number. Copies of these items are available on request from the user's local library, either for retention or for loan.
3. **Subject index.** As shown in figure 4, this is an alphabetic index based on descriptive terms assigned to each item of the bulletin. Each entry includes the title, document number, and page number of the bulletin where the abstract may be found.

Concurrently with the issuance of each of the three bulletins we distribute 16-mm microfilm cartridges or reels to our 26 library locations throughout the United States and foreign locations. We place on microfilm all IBM-originated documents and those non-IBM documents whose reproduction is permitted by copyright, and material that is releasable to foreign nationals. Documents are filmed in their entirety and are available to the user for his viewing and immediate reproduction of selected pages. Utilization of the microfilm at the various library locations reduces somewhat the requests for document hard copy.

The techniques and programs used by the IBM Technical Information Retrieval Center are not entirely new. We believe that the significance of our approach lies in the use of normal-text abstract searching and in the manipulation of a single input to produce a multiplicity of results, covering a broad subject area and serving a large diversified body of users. Although our efforts today are seemingly directed solely toward our user and the selection of input for his needs, we have not lost sight of the future and the fact that continuing study, research, and development must be made for the improvement of our Center's activities.

Mrs. A. S. Williams
Douglas Aircraft Company, Inc.

Historical Development and Present Status—

Douglas Aircraft Company Computerized Library Program

In early 1958, when responsibility for the Missiles Library of the Douglas Aircraft Company was transferred to its present section, the "under new management" sign was barely dry when the newly cognizant supervisory personnel were made acutely aware of the rapid growth rate of the Library's holdings. With the tremendous influx of technical documentation came the realization that traditional library practices were inadequate to cope with the problem. Accordingly a decision was made to index the library's holdings of technical reports by the "uniterm" method of cataloging (see item 3).[1] This method's potential for subsequent mechanized retrieval made it especially attractive. In January 1959 efforts were initiated to study the feasibility of expanding the manual uniterm system to a computerized document retrieval system. Computer management keenly interested in documentation problems and a librarian willing to forego traditional practices proved to be a fortunate combination. During the study and analysis phase, investigation was made of available retrieval systems, including those in being and those still planned. As the study progressed, advice was solicited from knowledgeable visitors to the Los Angeles area, including those of the stature of the late Pete Luhn of IBM and Dr. Mortimer Taube and All Kreithen of Documentation, Incorporated. These efforts culminated in May 1960 in a proposal (see item 1) recommending use of available time on an IBM 704 computer Rather than being an idealistic and theoretical treatise, the proposal aimed to justify the research and development effort required to develop a computer program

[1] This and similar references are found in the bibliography, page 54.

as a specific library aid and as a practical cost-saving device for library management.

Development of the System

Development commenced shortly after submittal and acceptance of the proposal. In November 1960 the program was expanded to include thesaurus, or dictionary, input from all divisions of the corporation and participation in the retrieval program by all company libraries. This move accounts for the wide range of subject matter encompassed by the system.

As initially planned the program was to have been implemented in three phases: a bibliographic subsystem whose by-products would materially reduce library clerical functions; a thesaurus, or dictionary, subsystem; and a retrospective literature search subsystem. With the advent of Automatic Selective Dissemination of Information (ASDI) (see item 5), it was evident that ASDI should precede the literature search phase. ASDI provided a more immediate payoff than the retrospective search, which requires a large inventory, because ASDI requires only a certain volume of input to be successful. The two initial phases, the bibliographic subsystem and the thesaurus subsystem, were developed concurrently and have been operational, for the most part, since May 1961. Development of the ASDI phase was partially a concurrent effort. It became operational about a year later. The literature search subsystem was developed last and has been operational since mid-1963.

Equipment Used and Products Produced

Let us now consider the system in more detail. Normally, input to the system is generated by copying a worksheet, compiled by a cataloger, on a Friden Flexowriter, which simultaneously produces paper tape and a hard copy for proofreading and interim usage. Originally, the punched paper tape from the Flexowriter was converted to punched cards by a Systematics Model C-750 Converter as an interim step to magnetic tape. This was replaced by the present method of producing magnetic tape directly from punched paper tape by means of a Control Data Corporation 160A computer. Initially, some problems of priority conflict, tape density, and tape quality were encountered because the conversion equip-

ment is under cognizance of one department and the computing equipment is under another. Establishment of good communication channels and close coordination solved these problems.

The computer originally considered was an IBM 704; however, during the developmental phase the IBM 704 was replaced by an IBM 709. Currently, IBM 7094 and 1401 equipment are utilized. (See fig. 1.) The 1401 is used purely as an input/output device

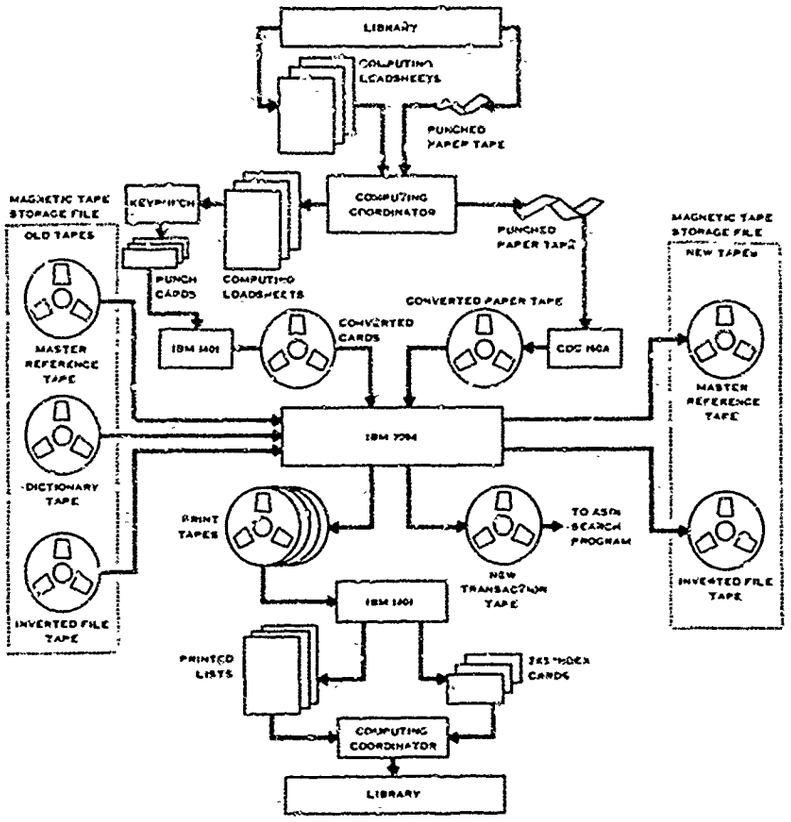


Figure 1

for the 7094. Punched cards are converted to magnetic tape by the 1401, and, in turn, magnetic tapes from the 7094 are converted to printed listings and cards by the 1401. A typical library input usually has about 500 additions to the master reference file every two weeks. The additions to the system are input by means

of the punched paper tape; however, corrections to the descriptors or uniterms are made by means of punched cards. The bibliographic program takes about 45 minutes to run on the 7094 every two weeks. Output consists of five magnetic tapes, four of which are designed as print tapes. One magnetic tape is generated with the newest acquisitions and is immediately input to the ASDI program. Of the four print tapes, one prints sets of 3 x 5 cards with information permuted and then sorted by accession number, author, title, and corporate source. Another tape prints out a two-part accessions list. A third tape generates a list of typing errors, a list of undefined descriptors, a list of accession numbers which are not on magnetic tape, a list of the number of 3 x 5 cards, by category, printed by the program, and the total of accession numbers currently on magnetic tape. The fourth tape prints a list of accession numbers which were added or replaced, a list of accession numbers on the magnetic tape which have incorrect descriptors, and a list of accession numbers not changed as requested together with reasons why they were not corrected. All this printing takes about one and one-half to two hours on the 1401 every two weeks.

Bibliographic Data Subsystem

In addition to the bibliographic data noted above, special information is generated upon request. A printout of document holdings by publication date assists in weeding our obsolete holdings. Regrading of documents with military security classifications is easily accomplished by citing both old and new classifications on a computer loadsheet. A set of cards reflecting the new classification is generated and substituted for the old set. If required, the computer can generate a listing of the total number of documents under each security classification.

For publication purposes, the accession list printout, which comes from one of the 7094 magnetic tapes mentioned above, is microfilmed, and then the microfilm is input to a Xerox Copyflo which produces offset masters for lithographic reproduction. In the near future this process will be abbreviated. The 7094 tape will go to a SC-4020 high-speed computer-recorder which will produce the required microfilm directly. Elimination of manual microfilming will greatly accelerate the availability of the data.

Another important product are the updated 5 x 8 uniterm subject cards used for manual literature searches. These are printed on request. Machine-generated cards have eliminated the long-hand or typewritten posting of new accession numbers to the subject cards.

The bibliographic data subsystem also produces statistical data. Typical data are listings of the number of documents produced in a given year, the average number of descriptors per document, the number of documents with abstracts, the number of indexed documents produced by Douglas, and a list of the accession numbers of those documents.

Other statistics are obtainable for assisting computing personnel to monitor the system with minimum effort. Some examples are tape length and length of record, checks on dictionary code-number assignment, frequency of additions or deletions to the dictionary, and checks on the updating of the inverted file.

Error lists are an important by-product of the system and are of sufficient interest to describe in detail. The "list of input errors" notes errors in the format of data entries, errors in accession numbers, errors of too many or too few fields (for material added or omitted), and so on. The "gap list" shows which accession numbers are not on the master tape. If there is a discrepancy, an entry may have been omitted, or a machine error may have occurred. The "replacement list" is a list of accession numbers which have replaced other entries on the tape. Entries are occasionally replaced to record additional information on the index cards or the tape record; for example, to show a superseded document. This list is used to verify the acceptance of the most recent entry and causes the previous entry with the same accession number to be replaced. A record of the number of entries on the old master tape and on the updated reference tape assures the Library that the correct tape was used and that no entries were lost during updating. The "term correction list" contains undefined uniterms and the accession numbers on which they appear, plus results from screening the bibligraphic input against the dictionary tape. All accession numbers appearing on this printout are flagged on the master reference tape and cannot be used for retrieval until the error is corrected. Any descriptor not in the dictionary is in error and is

printed in alphabetical order in the incorrect format with its accession number. If the descriptor is misspelled, it is simply corrected on the printout, without retyping. If, however, the word is a valid indexing term that for some reason does not appear in the dictionary, it must first be entered into the dictionary through the dictionary subsystem and then entered through the error sheet to remove the flag from the entry. The entire system, like all good systems, is designed for protection against what can go wrong, against the worst possible cases. For example, due to training of personnel (during one analysis period seven catalogers and eight Flexowriter operators were trained), the input error rate grew to 19.5 percent, but it dropped shortly thereafter to 15 percent and currently is less than 5 percent.

As an additional feature the program has the capability to store and print abstracts of the documents indexed. This capability is not being utilized at present because we feel that the list of descriptors adequately describes the document for our purposes.

Thesaurus Subsystem

The thesaurus subsytem was developed on the basis of documents already indexed by the three division libraries, and with reference to ASTIA, SAMPE, and AEC vocabularies. It might have been helpful to consult subject specialists or to use an established vocabulary, but this was not possible. For example, the ASTIA thesaurus was not pubilshed until more than a year after we created our basic dictionary. New terms are input to the dictionary after thorough researching of published vocabularies, consultation with user subject-specialists, and agreement by catalogers of participating company libraries.

Company-wide utilization of a single dictionary has resulted in a listing of over 11,000 terms, covering research, engineering, testing, and production, primarily in such areas as airframes, missiles and space vehicles, and encompassing such varied disciplines as astrophysics, biomedical and life sciences, electronics, flight mechanics, cryogenics, propulsion, structures, and so forth. The rate of growth of the dictionary has appreciably decreased in the past year. Obviously, the value of the entire system is dependent upon the contents and use of the dictionary.

Our approach to cataloging could be considered a compromise. Cataloging in depth with Ph.D. specialists would be expensive, but retrieval would be fairly inexpensive. Conversely, cataloging by clerks with only a high school background would be inexpensive, but this would necessitate a more complex and expensive retrieval system. Our catalogers, while not Ph.D.'s, do have degrees, some in library science. Clerical help is provided to them in order that the catalogers devote as little time as possible to clerical duties. Current research by others may prove that this approach is not the best, but it appears to be working well.

Although the "single uniterm" concept had been the goal, development of the dictionary as a consolidated vocabulary for all company division libraries disclosed the need for compromise and deviation. In our present usage "uniterm," "term," and "descriptor" are interchangeable. The vocabulary reflects "binding" of terms, acronyms, use of "scope notes" for semantic clarity, multiple spellings, cross references, and the inclusion of FC (frequency coordinated) terms. The program was expanded to give computer-generated postings of FC terms as a usage criterion for binding terms into one concept. Bound terms measurably decrease time spent in manual searching but present a hazard to thoroughness in either manual of computer searches if the break-off point of postings from single terms is not immediately reflected in the system. Computer-structured generic relationships have been used only for company-produced products. The slow progress that has been made to date in developing a structured vocabulary as a cataloger's working tool has proved that time, specialization, research, and difference of opinion are going to be the contributing factors in determining the feasibility of adding this feature to our dictionary program. An experimental modified-link system was used by one of the division libraries at the beginning of the program; however, the time spent in accurately linking descriptions seemed to be a more significant problem than "false drops" that have occurred during the several years of program operation. Our solution has been to use retrieval specialists who keep in close contact with the user's need through interviews and who make periodic checks of retrieval statistics. Also, we use catalogers who double at times as reference librarians.

Upon request, the dictionary tape can be used to print the dictionary, check the validity of uniterms on index cards before they are added to the master tape, and assigned unique code numbers to uniterms for use by the retrieval program when searching the tape files. These code numbers are internal to the computer and are not used as input or reflected in the output of the computer program. As a by-product the program also prints the frequency of use of each indexing term by each of the participating libraries (fig. 2). Descriptors are stored in an inverted file. If a descriptor is

2/13/64	PAGE 219		
HOMEOSTASIS	5	0	3
HOMING	33	1	19
HOMOGENEOUS E.G. HOMOGENEOUS REACTORS	39	8	60
HOMOGENIZATION	0	3	3
HOMOGRAPHIC	1	0	1
HOMOPOLAR	2	0	1
HONEST JOHN =SEE ALSO XM-50, M-W1, ROCKETS 762MM, XM-37	108	1	2
HONEYCOMB, HONEYCOMB SEE ALSO ATRCOMB, SANDWICH	102	107	102
HONING SEE ALSO LAPPING, SUPERFINISHING	0	1	0
HOOKS, HOOK	1	1	12
HOPF WIENER- SEE WIENER-HOPF			
HORIZON	11	0	28
HORIZON (PROJECT)	0	0	3
HORIZONTAL	23	12	107
HORIZONTAL STABILIZERS, HORIZONTAL STABILIZER	0	0	0
HORIZONTAL TAILS SEE HORIZONTAL STABILIZERS			
HORIZONTAL TAKE-OFF, HORIZONTAL LANDING SEE HTOHL			
HORMONE, ADRENOCORTICOTROPIC SEE ACTH			
HORMONE, ANTIDIURETIC SEE VASOPRESSIN			
HORMONES, HORMONE	14	2	8
HORNS, HORN	1	0	26

Figure 2.

input before it is established as a valid term, the dictionary tape will reject that entry as input and flag it for listing on the term-correction list.

The combining of terms already a part of the dictionary is accomplished by adding them to the dictionary loadsheet and requesting that they be added to previously indexed material through the "correction of uniterms" loadsheet. The same method is used to delete terms. As a precaution against deleting a term inadvertently, the dictionary program will not accept a deletion while there are still items posted on the term, but it will print a list of accession numbers which must have the term removed. These built-in safeguards are well worth the effort required for their development and are highly recommended to designers of retrieval systems. The dictionary tape is the hub of the entire program; it ties the bibliographic subsystem to the retrieval subsystem, and upon its accuracy rests the success or failure of retrieval.

As an extra feature each document has affixed to it a statement inviting the scientist/engineering-user to analyze the assigned descriptors and suggest corrections or additions. Also, as a company policy, authors of company reports are to assign descriptors, in collaboration with the library catalogers, at the time the report is written. Our feeling is that if all originators of documents met a similar requirement, eventually (discounting the present problems caused by variances in vocabularies or dictionaries) cataloging efforts throughout the country could be reduced; in effect, recipients would receive partially "precataloged" documents.

Automatic Selective Dissemination of Information (ASDI) SUBSYSTEM

Operational since April 1962, the ASDI subsystem was given priority in order to increase the payoff of the system, by making users more rapidly aware of newly acquired documents pertinent to their areas of interest. Subject profiles for individuals and Douglas organizational units are constructed of terms in the mechanized program dictionary after users are interviewed by a retrieval specialist in the Library. A tape produced from loadsheet input is matched against the regular semimonthly library

accessions input. Output is in the form of a library index card coupled to a reader response card (fig. 3).

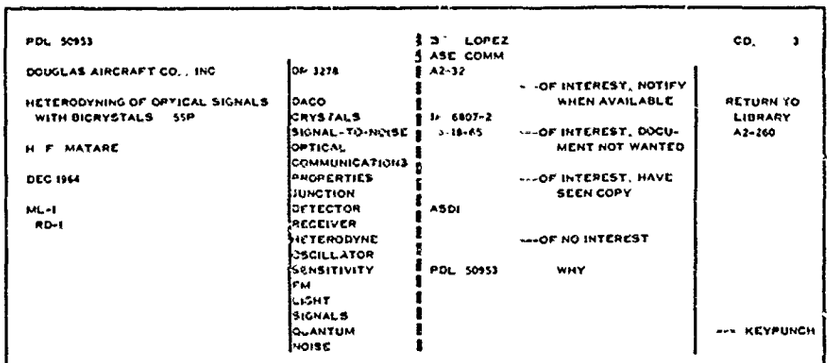


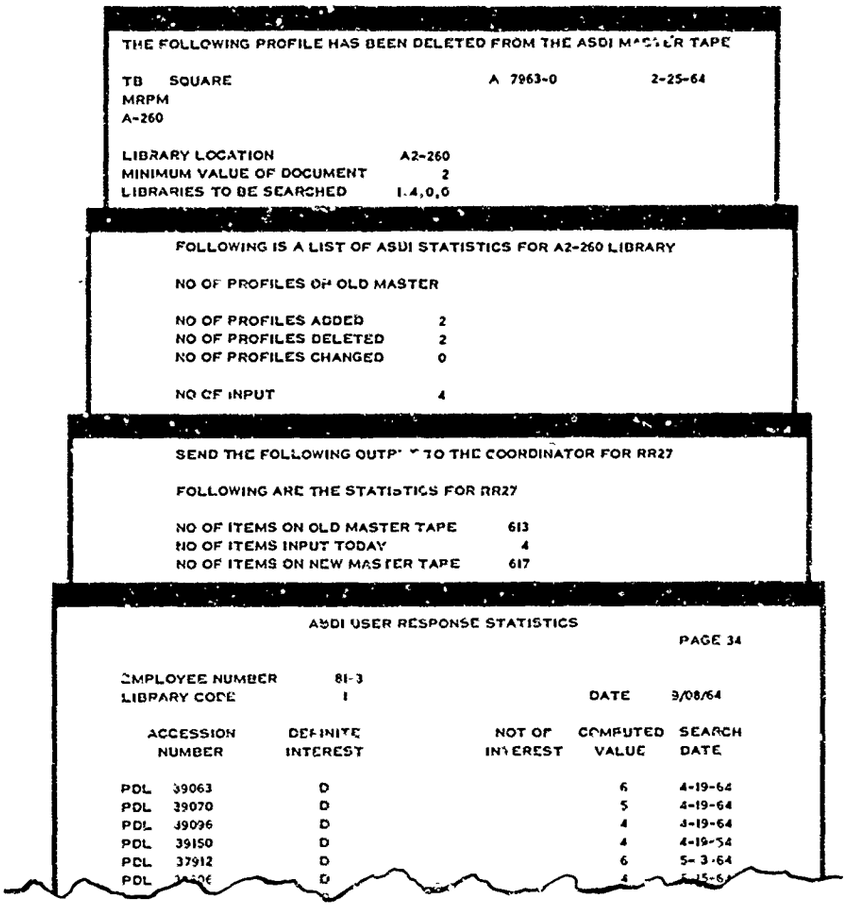
Figure 3

There are now about 125 profiles of individuals and Douglas organizational units in the ASDI system. Analysis at an early stage indicated about a 50 to 60 percent "hit" ratio, while the adjusted or "polished" profiles gave positive responses with ratios as high as 91 percent. This indicated several things, primarily that the search program is flexible enough for the vocabulary to be tailored to the individual; that the unstructured descriptor vacabulary can describe document content adequately; and that care in initially establishing the search question is extremely important. Statistical output (fig. 4) assists in monitoring the ASDI program.

The ASDI program is run immediately after the bibilographic data program. The total running time on the 7094 takes about 15 minutes every two weeks, with printing and assembling taking about an hour. (These times refer to a typical division library, not to the total corporate effort.)

Literature Search Subsystem

Input to the computer for a retrospective literature search is very similar to that for the ASDI subsystem. The basic design of the loadsheets, tape format, search logic, and printed output for both programs is much the same. The literature search profile is used for a one-time input in searching all material on file in the mechanized system. The user may specify the number of retrievals desired, indicate which division library tapes are to be searched,



THE FOLLOWING PROFILE HAS BEEN DELETED FROM THE ASDI MASTER TAPE

TB SQUARE A 7963-0 2-25-64
MRPM
A-260

LIBRARY LOCATION A2-260
MINIMUM VALUE OF DOCUMENT 2
LIBRARIES TO BE SEARCHED 1,4,0,6

FOLLOWING IS A LIST OF ASDI STATISTICS FOR A2-260 LIBRARY

NO OF PROFILES ON OLD MASTER

NO OF PROFILES ADDED 2
NO OF PROFILES DELETED 2
NO OF PROFILES CHANGED 0

NO OF INPUT 4

SEND THE FOLLOWING OUTPUT TO THE COORDINATOR FOR RR27

FOLLOWING ARE THE STATISTICS FOR RR27

NO OF ITEMS ON OLD MASTER TAPE 613
NO OF ITEMS INPUT TODAY 4
NO OF ITEMS ON NEW MASTER TAPE 617

ASDI USER RESPONSE STATISTICS

PAGE 34

EMPLOYEE NUMBER 81-3
LIBRARY CODE 1 DATE 3/08/64

ACCESSION NUMBER	DEFINITE INTEREST	NOT OF INTEREST	COMPUTED VALUE	SEARCH DATE
PDL 39063	D		6	4-19-64
PDL 39070	D		5	4-19-64
PDL 39096	D		4	4-19-64
PDL 39150	D		4	4-19-64
PDL 37912	D		6	5- 3-64
PDL [illegible]	D		4	[illegible]

Figure 4

establish date parameters, and indicate the form of output, as cards or lists or both. The requester's search profile of dictionary uniterms is tailored to his requirement—a broad and comprehensive search of direct and related subjects, or a very narrow and specific search The choice of descriptors is influenced by the frequency-of-use statistics listed in the dictionary. Each search profile input has a minimum value for any document the search is to retrieve. The document value is determined by the sum of the weights of the descriptors common to the search profile and to the document in question. The weight value given to a single descriptor is determined by the importance of the descriptor to the user's requests

and by the frequency of its use in the system. The requester may further limit the search by listing descriptors which would make the retrieval *undesirable* to his particular area of interest.

It may be well to point out here the danger of a library to tend toward complacency—initially, a user is so appreciative of receiving *any* information that the search librarian has to be super-critical in analyzing the retrieved information to ensure that the search has been thorough.

The literature search program can be run any time the library has need for it; no set schedule is established, although it currently is run about twice a week. The MSSD Engineering Library is averaging 50 searches a month. Up to 100 searches can be made during any one computer run. Although the program logic is very similar to the ASDI search, the running time is usually longer due to its searching each library reference tape completely. The 7094 time averages 25 minutes; however, printing and assembly time can vary anywhere from 30 minutes to 6 hours, based on the quantity of retrieved information.

Possible Extensions and Improvements

At the present time over 97,000 documents (primarily classified and unclassified technical reports) from all Douglas libraries are indexed and on tape. Excluded are books, periodicals, journals, and specifications for which other retrieval methods exist.

It was soon realized that ensuring the most efficient use of a successful and operating computerized retrieval program required computerization of as many other library activities as possible. Programs such as one for mechanized routing of library accession bulletins (DDC's *TAB* and NASA's *STAR*) were implemented. A statistical listing from a 1401 computer program assists in reducing the clerical workload in the periodical area. The list gives the number of periodical subscriptions, frequency of distribution, date of renewal, date of binding, and general subject area.

At present, studies are being made of mechanized or computerized circulation systems. An acceptable system must include a quick and easy method of input, short turnover time in the Computing Systems Operations Department, and it must aid in evaluating and deleting entries already on the master reference tape. NASA's

computerized system and tape output has been followed with interest since first it was announced, and periodic studies have been made to develop its compatibility with the Douglas program. The recent announcement that NASA's files have been reorganized to linear sequence and that 1401 and 7094 programs will soon be available will expedite completion of the Douglas effort. Once in effect, the scope of in-house literature searches will increase proportionally with the number of NASA microforms on file. There are, already, over 100,000 microforms in the Library's holdings. The Douglas program will also be readily adaptable to any other organization-issued magnetic tapes and microfilm.

With the advent of new computing equipment and a review of the entire mechanized program, it is hoped that computerized acquisitions and receivals will be possible. Computer detection of duplicate material at the acquisitions level would be coordinated with all company division libraries. Automatically generated follow-up notices on ordered material would also be a part of this program. Thought has been given to eliminating the necessity of using punched paper tape. Optical scanning has been investigated, but in all likelihood System 360 auxiliary equipment would be used. Also being studied is the use of book-type (index) printout format in lieu of 3 x 5 cards for less-active accession areas.

In the computing area, improvement currently contemplated will be aimed at greater efficiency in the software, i.e., in the subsystems, sorting routines, and so forth. The initial program was based on what was then considered average input; however, the input, about 1,800 documents a month in spite of careful screening, is considerably over expectations. The computing logic was based on information available in 1959 and has proven to be very good. Today's computing people are impressed by the foresight shown by their predecessors.

One area which hasn't received adequate attention is that of formally documenting the logic. This has made it difficult for personnel unfamiliar with the program to pick it up. An early decision was that the program would be peculiar to Douglas, so it was set up the "Douglas" way. This is proving to be a handicap when it comes to encouraging others to use the program. When the program is rewritten for IBM System 360 equipment, it will

be converted to a "universal" computer language, such as COBOL or NPL, thus making it easier for others to adapt to their requirements.

Conclusion

In retrospect we are gratified that the present system bears so much resemblance to the theoretical program first proposed. That in itself is a tribute to the soundness of the basic system. The developmental period required a lot of effort and was even frustrating at times, as is often the case when theory is transformed into practice. The changes made as the program evolved were definitely beneficial, and their description is passed on with the hope that they may be of value to others currently in a developmental phase. We anticipate no major upheavals in the program due to contemplated changes. The program has weathered, without strain, not only the transition for 704 to 7090 and 7094 equipment, but the expansion from one division library to several; so, we are satisfied that basically it is a system ideally suited for our requirements.

BIBLIOGRAPHY

1. Bunnow, L. R. *Study of and Proposal for a Mechanized Information Retrieval System.* Report No. SM-37418. Douglas Aircraft Company, Inc., May 1960.
2. Bunnow, L. R. and Koriagin, G. W. *Mechanized Information Retrieval System, Status Report.* Report No. SM-39167. Douglas Aircraft Company, Inc., January 1962.
3. Documentation, Incorporated. *The Uniterm System of Indexing, Operational Manual,* 1955.
4. Douglas Aircraft Company, Inc. *Douglas Mechanized Library Program Panel Presentation.* Report No. G-36444, April 10, 1964.
5. Luhn, H. P. *Selective Dissemination of New Scientific Information with the Aid of Electronic Processing Equipment.* Yorktown Heights, N.Y.: International Business Machines Corporation, Advanced Systems Development Division, November 30, 1959.

PANEL DISCUSSION SUMMARIES

PANEL SUMMARY

Bernard K. Dennis
Battelle Memorial Institute

Indexing and Classification

The eleven members of the Indexing and Classification panel came from Government and non-government organizations and represented a wide range of interests, training, and experience in the science information field. The semiautomated systems they represented varied from a newly initiated system, with only a few thousand documents in file, to a large, general-purpose, computer-based system providing selective dissemination to several hundred users at locations throughout the nation.

Perhaps the most striking feature of the group was the relatively calm, almost cautious, manner in which various aspects of indexing or classification were approached. Only a few years ago, this kind of discussion would have been dominated by hotly pressed claims and counterclaims by zealous proponents of the latest cure-alls for the world's indexing or classification ills. It was heartening indeed to observe the degree of objective openmindedness that permeated the group's deliberations.

Although many issues in indexing and classification were dealt with by the group during its day-and-a-half discussion, topics fell generally within the following areas:

- Classification
  - COSATI [1] subject groups and announcement media
- Indexing
  - Indexing approaches used in represented systems
  - Advantages vs. costs of sophistication
  - Author indexing
  - Selection and training of personnel
  - Quality and quality control
- Standardization
  - Levels of opportunity
  - Human and machine indexing

Classification

There was little interest in classification beyond its use in announcement media. Main interest was focused on the twenty-two COSATI subject groups established for achieving uniformity in Government media announcing the availability of scientific and technical documents. Although it was the consensus view that uniformity would be desirable, serious concern was expressed over the practicability of the COSATI scheme. Greatest concern stemmed from the fear that many of the users for whom the announcement tools are intended may not readily find their areas represented. There could be a significant loss in utility as a result of the attempt to force an unwelcome change in habit patterns upon the users, who very probably would become unresponsive and thereby defeat the purpose of the announcement media. The consensus of the Indexing and Classification panel was that COSATI should proceed more cautiously than is apparently planned and should take into fullest possible account the affect of COSATI direction on the information requirements and habits of the user groups for whom the various announcement media are intended.

Indexing

With two exceptions, the indexing approaches of the systems represented by participants in this discussion were based on coordinate indexing and computer-manipulated inverted indexes. One system used a computer-produced, keyword-in-context index with provision for cumulation. Another operational system, for storing legal documents, was based on complete text natural-language processing by computer. The participant concerned with this system indicated that user acceptance has been satisfactory. He contended that the approach may be generalized beyond legal documents, but he agreed that economic justification would depend upon the situation.

---

[1] COSATI is the acronym for the Committee on Scientific and Technical Information, Office of Science and Technology of the Executive Office of the President.—Ed.

Sophistication—advantages vs. costs

The subject of indexing methodology was pursued somewhat lethargically by the group (or so it seemed to the chairman). With one exception, no one appeared to feel strongly committed to a particular approach. The consensus seemed to be that the inverted coordinate index lends itself well to machine manipulation and is probably the most economical basis today for effective semiautomated document retrieval. One system represented at the workshop was based on a hierarchically structured inverted coordinate index manipulated by computer. This system, in operation at the Science Information Exchange at the Smithsonian Institution,[2] probably represented the highest order of sophistication of any of the conventional systems discussed.

None of the system operators at the workshop used a vocabulary or index-term control device other than a thesaurus. There was general concern that other devices might reduce recall, even though they achieved an element of standardization and improved the relevance of retrieved material. The benefits of role indicators in specific situations were acknowledged; however, caution was expressed regarding their general applicability. Considerable emphasis was placed on the economic and intellectual requirements dictated by their reliable application.

Author Indexing. Regarding the possibility of utilizing author indexing as direct input to a system, it was unanimously agreed that such practice is not likely to be feasible. It was felt that even though potential authors were taught to index, the infrequency of their need to do so would not permit them to maintain an adequate degree of proficiency. However, it was generally felt that preliminary indexing by the author, followed by professional indexing, would tend to be significantly beneficial, by assuring that the main intent of material was reflected in the total indexing.

---

[2] A description of the S.I.E. system is on page 62.

Personnel Selection and Training

The highly subjective nature of manual indexing was made apparent during the discussion of the selection and training of indexers. The prime requisite for a candidate indexer was felt to be subject knowledge. Second was personality. Panel members hotly refuted the notion, often been expressed in some quarters, that indexing requires no subject-matter expertise in the field being indexed. The group could give but few guidelines for training indexers. The process apparently depends heavily on intuition and thorough early review and supervision. A particular system's thesaurus and indexing ground rules appear to be the main teaching aids.

Quality and Quality Control

Assurance of indexing quality appears to be a highly subjective process, with close supervision and review by experienced indexers being the main approach. Mechanical editing, insofar as spelling, word form, symbols, etc., are concerned, is feasible; but assuring that chosen index terms really convey the message content of the material being indexed is a subjective matter. If a high-grade thesaurus exists for the system, mechanical posting will assure the same quality reflected in index posting. The group readily admitted that, in the final analysis, quality of system output is a function of the quality of system input, which is largely an immeasurable characteristic.

Standardization

One of the areas receiving major attention by the group was that of standardization. It was agreed that although standardization is of key significance for purposes of system efficiency and economy, it is a prerequisite for efficient and economic interfaces among systems. However, the effects upon the users of achieving standardization and/or compatibility may be such that a system's purpose may suffer to an unacceptable extent.

Levels of opportunity

The group agreed that there are four levels of opportunity for achieving standardization and/or compatibility. Ranked in order

of their practical achievement they are as follows:

1. Descriptive cataloging
2. Classification for announcement
3. Published subject indexes
4. Indexing for document retrieval

It was readily agreed that the attainment of standardization in descriptive cataloging poses no great problems and that it could and should be achieved on a widespread scale. The feeling regarding standardization of classification for announcement or distribution purposes was indicated by the group's concern over the COSATI scheme. Caution is required due to the impact of decisions upon the users. The same need for caution applies in attempting to standardize printed subject indexes. Users' habit patterns are of paramount concern.

Regarding deep indexing for subject retrieval of documents, the group felt that possibilities for standardization within a system and compatibility among systems appear remote for a number of reasons, the chief one possibly being the difficulty of coping with the human element.

Human and machine indexing

It was the general feeling of the group that, for the foreseeable future, manual indexing for document retrieval will be a major cost element in system operation. Little hope was held for significant improvement in quality and reliability. It was considered that only less-than-optimum benefits can be gained through mechanizing index searching or manipulation as long as system input depends upon today's indexing capability. In the group's opinion, it appears highly unlikely that the intellectual function of selecting significant retrieval terms can be performed with adequate precision solely by computer, at least within the near future. Also, there was concern about the economics of natural-language processing of complete text by computer, as well as about retrieval recall and relevancy in a completely automated system.

Conclusion

At the present time, there remain limitations in the arts of indexing and classification which apparently prevent realization of optimum results at commensurate cost. These limitations include

or are related to:

- Subjectivity of the human indexer
- Size and rate of growth of the document base
- Purpose and patterns of use of systems, tools, and media
- Available resources

It appears that some levels of standardization and/or compatibility in selected areas would help reduce the effects of human subjectivity and would promote efficiency and economy. Standardization in descriptive cataloging should be relatively easy to achieve. A significant degree of standardization and/or compatibility in announcement classifications and printed indexes may be accomplished with trade-offs acceptable to the users.

A concerted effort to achieve agreement in indexing methodology and in standardization and/or compatibility in deep indexing could have a profoundly beneficial impact on the utilization and allocation of resources available to the science information field. However, what this might mean for the user is not clear.

Technological advances in information processing have already made possible, if not operationally feasible, the automatic indexing of natural language and retrieval from complete text. Thus, it appears that there may be a possibility of resolving some of the difficulties inherent in manual indexing. At the very least, a capability for consistency and measurement would accrue. The Indexing and Classification panel therefore made the following recommendation:

> Recognizing the need for improved understanding of and refinements to manual indexing and classification methodology, we recommend that research and experimentation in machine processing of natural text be pressed from the standpoints of utility to the user, of requirements for software and hardware, and of the impact of such capability throughout a total information system.

HIERARCHIAL CLASSIFICATION AT S.I.E.

INDEX	COMPUTER CODES				REFERENCE POINTS
MEDICINE					
370					BLOOD
	01				BASIC
		120			PLASMA
	19				BLOOD PICTURE IN DISEASE
		600			PLASMA
PHYSICS					
7800					PLASMA - PHYSICS
	72				PLASMA - PRODUCTION
	76				TYPES OF PLASMA (20 SUBCATEGORIES)
7850					SOLID STATE
	73				RESONANCE IN SOLIDS
		670			PLASMA
EARTH SCIENCES					
3685					MINERAL INDEX
	80				SILICATES, TECTOSILICATES
		759			SiO_2 GROUP
			68		QUARTZ-CRYPTOCRYSTALLINE
				632	PLASMA

This system represents the most logical approach to classifying research ideas and ongoing work, since the scientific mind is trained from its earliest days in this type of scheme.

To the keynoter's remark that there might be "too much emphasis on hardware" we should like to add the first recommendation of the President's Science Advisory Committee, that scientists and engineers must be urged "to commit themselves . . . to handling information with sophistication and meaning, not merely mechanically." [1]

We at S.I.E. would like to recommend that information centers which are to serve scientists and engineers consider the following in connection with indexing and classification:

- Hiring indexers, analysts, and other subject-matter specialists who are highly educated in their respective fields.
- Giving these specialists appropriate recognition in status and pay.
- Making available to these specialists improved hardware which would speedily store, process, coordinate, or otherwise manipulate words and concepts.
- Minimizing the basic problem of semantics by allowing the users to make inquiries in their own scientific and technical terminology and by not forcing them to depend on thesauri and other "authorized" word lists.

We are somewhat doubtful that present indexing of interdisciplinary areas will satisfy the discriminating scientist and engineer to the point of complete confidence in a system. We go along with others who advocate more research in hardware and word manipulation, but we wish to emphasize that these activities are only a part of the whole picture of indexing and classification. To summarize, we believe in the man behind the machine rather than in the machine as such.

---

[1] Found in *Science, Government and Information*, the "Weinberg Report."

PANEL SUMMARY

Herbert Rehbock
Defense Documentation Center

Abstracting and Extracting

The topics discussed by the panel included the definition of abstracting and extracting; the purpose, style, and authorship of abstracts; the criteria for abstract preparation; the qualifications of abstractors; and abstracting by computer.

Definition of Abstracting and Extracting

Definitions of abstracting and extracting were developed as a point of departure for the discussions. Members described early abstracting efforts and noted that these had the purpose of making a scientist read a paper and write properly. *Abstracting* was defined as the summarizing of a document or published material briefly and accurately. An abstract is prepared from the substance of a report and is reworked into a summary by an abstractor. It is utilized to indicate the discipline orientation and can serve as a proxy or substitute for the document. *Extracting* is the taking of data and information directly from a document in the exact words of the original material. An extract is primarily intended for the use of subject specialists.

Purpose of Abstracts

Having agreed upon definitions of abstracting and extracting, the group deliberated on the purpose of abstracts and the reasons for their preparation. There was considerable agreement as to the purposes of an abstract, which can be summarized as follows: (1) to let the reader know that the report exists, (2) to permit the reader to select the report for perusal or eliminate it from further consideration, and (3) to serve as a substitute for a report. When foreign language material is being reviewed, the abstract should be in the reader's language.

Important parameters affecting abstracting include completeness, readability, ease of indexing when the actual document is not

available, accuracy and absence of bias in presentation, consistency in bibliographic entries, and practicability and obtainability. Several members emphasized that in their experience only the easier documents are abstracted and that the more difficult material is treated in a very superficial manner. This applies particularly to documents containing the proceedings of symposia, professional meetings and conferences. It was recognized that material of this type cannot be covered by a single abstract in any adequate manner. The question was raised why these documents are not abstracted so that each paper presented at a symposium or conference would carry its individual useful abstract. In this regard these was concern expressed about abstracting the briefs of proceedings which are merely summaries of papers not available for duplication.

Types of Abstracts

The next discussion centered around the style and length of abstracts, which are exemplified by (1) the telegraphic type, (2) the indicative type, and (3) the informative type. It was decided that telegraphic abstracts have been used primarily for indexing purposes or as an adjunct to the conventional abstract. It was also noted that this type of abstract is becoming more important in computer applications. One style of telegraphic abstract that was discussed employed symbolic language to present only the purpose, procedure, and methods reflected in the report. This style has found acceptance in the medical sciences, and it is probably true that it will suffice for all the life sciences.

The indicative abstract, sometimes referred to as a descriptive abstract, provides only the necessary and relevant factors about what was done; it does not report the findings. The indicative abstract lets the searcher know what he will read if he goes to the original document.

The most desirable type of abstract is the informative type, which gives in a concise manner the details or what was reported and provides facts and findings. It should consist, whenever possible, of the purpose, method, results, and conclusions found in the report.

The beginning statement for an abstract should inform the

reader of the reasons for an investigation. The abstractor, however, should not merely re-emphasize the title of the report. If the title is very informative, the abstract can begin with a description of how the research was accomplished. The reader should be told what methods were employed, how it was done, with what material or data, and under what circumstances. Here the abstractor must weigh his words and be brief. What was learned from an investigation is probably the most important material to be included in an informative abstract. Often there are too many specific results for inclusion. To prepare a thoroughly informative abstract, the abstractor should present what was found to be *new* in the field under investigation. In the conclusions, he should include a statement as to what may be derived from the investigation, what it meant, and how it may be of value or interest to investigators in similar fields of endeavor.

Review of the application of informative abstracts to technical reports and scientific papers was followed by a discussion on the application of the informative abstracts to progress reports, evaluation and qualification test reports, and bibliographies.

Abstracting Criteria

The group next examined the selection criteria for abstracts—the standards that should be employed for the selection of significant data contained in the source documents, when to abstract and when not to abstract, how to prepare scope or content notes, and how to use brief annotations or amplification of titles. For a document in a specific discipline, the title and author often provide sufficient clues to the reader; but abstracts are necessary in interdisciplinary areas. Opinions were expressed about whether the abstract is a factor in the availability of a document. Several participants believed that if the document itself is readily available, abstracting can be kept to a minimum. In this case, a brief amplification of the title and/or a brief annotation is acceptable. The title can also be amplified through the use of key words, descriptors, and the like. An acceptable method of conveying to the reader the content of difficult-to-abstract books and symposium proceedings is to enumerate the content of such material. In the case of symposia, these enumerations should contain, as a minimum, the title of each paper and the name of its author or authors.

Modular Content Analysis

The coverage of a document as reflected by the modular content analysis concept was explained. Modular content analysis endeavors to prepare five different modules: (1) the title, (2) a brief annotation, (3) the indicative abstract, (4) the informative abstract, and (5) the critical review. In this concept the use of the title and the descriptive header-listing, supplemented by key words and descriptors (i.e., subject matter indicators), can suffice as tools for in-house preparation of accession lists. Title, annotations, and indicative abstracts can be utilized for announcement media in single-discipline publications. A combination of these with the addition of informative abstracts on a selective basis can serve the purpose of those announcement media which have an interdisciplinary character. The informative abstract is required for retrospective searches of computer or semiautomated data banks, as exemplified by the preparation of subject matter bibliographies. To state it in other terms, the value of abstracting must be judged in relationship to the reader's (user's) needs.

Author Abstracting vs. Centralized Abstracting

The group explored the question of who should prepare an abstract. There appears to be two primary sources for the generation of abstracts: (1) the author and/or the technical writers and editors within the originator's organization, and (2) a centralized abstracting service. For economic reasons and because of the tremendous amount of scientific and technical material generated, much of the abstracting function has been shifted to the author of a report. This was considered undesirable by some members of the working group, who felt that proper location of the abstracting function was in a central abstracting service. Apart from economics, the rationale for advocating author-prepared abstracts is based on the fact that the author is knowledgeable of the research he has performed, whereas a central abstracting service is not. The central abstracting service may inadvertently misinterpret the information and consequently misinform the reader of the abstract. On the other hand, author-prepared abstracts must guard against unjustified emphasis of the value of a report. In general, it can be said that the author should prepare

the abstract with a view toward the distribution of the information at a central location, and should bear in mind the needs of the respective scientific communities.

Central abstracting services have a requirement to modify the author-prepared abstract. This usually happens when the abstract is not suitable for the machineable records maintained by a central computer operation, or when it is necessary to eliminate stereotyped phraseology. A central abstracting service should modify an author-prepared abstract as little as possible. There appears to be no major objection to modification of format for standardized computer application.

Abstractor Qualifications

A brief discussion of the types of personnel employed in abstracting dealt with the use of professionals (i.e., scientific personnel) at the originator's location, and with the possibility of employing semiprofessional personnel in centralized documentation efforts. The qualifications for abstractors are related to the type of abstract desired and to the reliability that can be ensured with regard to the technical accuracy and validity of reporting. Accurate abstracts dealing with evaluation or critical reviews of research and technology require personnel with qualifications in particular areas or specialties. In experiments with modular content analysis, it was found that the critical review cannot be adequately prepared by the generally available abstractor, even though he may have the necessary basic scientific education. Because the need for abstracts varies from the highly informative and review types to the more descriptive or annotated types, the personnel required vary from professional to semiprofessional to editorial.

Automatic Abstracting

The group touched on the basic concepts involved in abstracting by computer and explored the subject of automatic abstracting. One case that was presented dealt with an existing ADP extracting procedure triggered by clue words in the text of documents. In automatic abstracting, a subset of sentences in a document is selected on the basis of statistical characteristics as being representative of the general content of the document. The difficulty

in automatic abstracting lies in bringing out the discriminating aspects or factors which make one document different from another on a similar subject. From the information gathered within this small study group, it seems that present effort is primarily directed toward extracting information from the actual text. Automatic abstracting appears to be a more vexing problem than was originally contemplated.

Machine language of abstracts poses a very real problem since many general-purpose printout devices do not provide the Greek alphabet, subscripts, superscripts, and mathematical symbols. Abstractors will have to continue to verbalize these symbols until new computer equipment with additional printing characters becomes available.

Conclusion

The members of the panel agreed that no one type of abstract is suitable for all purposes. The need for abstracts varies with the availability of the document, the requirement of the reader, and is conditioned by economic and personnel limitations. Different presentations must be prepared for different reader communities. The mode of announcement, i.e., announcement in local accession lists, discipline-oriented publications, or interdisciplinary publications, will determine the type of abstract each publication should utilize. Uppermost in the mind of the abstractor should be the fact that abstracting is a support to the ultimate user, which means that an abstract should be prepared so the reader can derive the greatest benefit from it.

I wish to thank the members of the panel on abstracting and extracting for their spirited discussions and valuable contributions to analysis of the subject at hand.

PANEL SUMMARY

William B. Hammond
Datatrol Corporation

Vocabulary Construction and Control

For the most part, the panel was composed of people with extensive experience in the topics that were discussed.[1] There were no radicals, no free indexers, no unitermers, and no proponents of Webster's *International*. The panel was unanimous in its belief that every facet of subject indexing requires continuing direction and control. There was near unanimity even in the details of such control. Thus, it was the consensus of the group that some form of subject indexing authority is essential to successful operation of an information system. It was also agreed that the authority list should include scope notes, cross references for posting, "see" references, and an indication of generic relationships among the authorized indexing terms. This display of generic relationships should reflect those generics actually encountered in the system and should guide the user to broader or narrower terms actually utilized by the indexers to discriminate among levels of specificity of the documentation being indexed.

Subject Lists

It was also the consensus of the group that a structured display of terms to some broader categories was required. Such a display would serve indexers as well as searchers. At this point, the group reviewed briefly two publications that were furnished by the Chairman: *Common Vocabulary Approaches for Government Scientific and Technical Information Systems and COSATI Subject Category List* (See items 2 and 3.) The former publication gives a brief description of the vocabulary lists of the major government agencies; authorities used for nomenclature; methods of introducing changes; and general considerations governing trade names, chemical nomenclature, and medical nomenclature.

The *COSATI Subject Category List*, issued by the Office of Science and Technology of the Executive Office of the President,

provides a subject category list for (1) announcement and distribution of scientific and technical reports issued or sponsored by the Executive agencies, and (2) management reporting. The list has a two-level arrangement consisting of 22 major fields, with a further subdivision of the fields into 178 groups; scope notes are included for each group.

Since the *COSATI Subject Category List* is rather recent, no member of the panel had yet had an opportunity to make a full appraisal of it. A few were making an effort to employ COSATI categories as a major subsumption scheme for displaying terms. It was mentioned that DDC was attempting to develop a similar display for the terms listed in its *Thesaurus of Descriptors.* However, DDC has found it necessary to add several additional categories to provide a meaningful display of its vocabulary.

Vocabulary Control

There was agreement on the need for some tool to provide vocabulary control, but the panel found that the term "thesaurus" stirred up disproportionate hostility. Only a half-hearted attempt was made to agree on a universal definition of an "Information System Thesaurus."

The panel also agreed that explicit instructions prescribing uniform procedures for applying subject indexing terms should be furnished each indexer. Furthermore, the panel felt that computers should be used to edit all indexing data processed into the system. At a minimum this edit should include automatic coding when numeric codes are substituted for alphabetic terms in a mechanized system file. If mechanized capability is available, it is also desirable that the edit should attempt to trace the human indexing patterns to determine that indexers are following the prescribed instructions. The computer should be employed to compile statistical information on the usage of individual terms, the frequent co-occurrence of indexing terms among the indexing data, and the frequent coordinations of terms in searching. The panel members have found these to be of value to searchers and computer operators and also of value for control of indexing (e.g., by examining very high or low postings).

---

[1] Comments by three panel members are found on pp. 74-76.

Machine File Organization

In addition to discussing the main topic of vocabulary construction, a rather lengthy discussion was held on machine file organization (serial vs. inverted files). Desirable and understandable features of different file operations were discussed. It was agreed that it was most desirable to examine the full citation available in the serial file format, rather than fragments of the citation provided by the inverted file, in determining the relevance of a given report to a search requirement. Some members, however, were employing inverted files because they believed (or their computer support facility believed) that certain economies were attained by employing this format in computer manipulation of their files or in computer-produced aids for manual access to data contained in the machine files. No firm conclusions were reached; the panel's discussion of file organization was primarily of an informative nature.

Recommendations

In general, the panel felt that workshops of this kind were very useful and productive and should be continued. Perhaps the topics to be covered by individual panels should be broadened, or individual members should be given an opportunity to participate in more than one panel–at least in our case this appeared desirable.

As Chairman, I am of the opinion that if subsequent workshops of this nature are scheduled, several papers should be presented on basic file organization–serial vs. inverted vs. random access (or "immediate" access) files.

BIBLIOGRAPHY

1. Wall, Eugene, *Information Retrieval Thesauri.* New York: Engineers Joint Council, November 1962.
2. Hammond, William, and Rosenborg, Staffan. *Common Vocabulary Approaches for the Government Scientific and Technical Information Systems.* Silver Spring. Md.: Datatrol Corporation, December 1963. (Datatrol Corporation. Technical Report IR-10, Contract NSF C-342). ASTIA Document AD 430 000.
3. Federal Council for Science and Technology. *COSATI Subject Category List.* Washington, D. C., December 1964. ASTIA Document AD 612 000. (Also available from CFSTI as PB 166 877).
4. Defense Documentation Center. *Guidelines for Using ASTIA Descriptors.* Cameron Station, Va.: Defense Documentation Center, March 1965 (reprinted).

COMMENTS

Harold B. Thomas,
Air Force Materials Laboratory

Vocabulary for Air Force Materials Information Retrieval System

Since no similar vocabulary was available, the vocabulary for the Air Force Materials Information Retrieval System was initiated by indexing the system's documents. Synonyms were placed under the same machine code from the start, and both roles and links were used in indexing. A thesaurus was assembled, and has been updated as changes occur.

As the result of an experiment, roles have been eliminated. Bound terms are used rather freely, and only one link is used except when necessary to prevent false coordinations. The use of many specific terms, especially in organic chemistry, had increased the vocabulary to between 45,000 and 50,000 terms—an excessive number under our circumstances. To improve this situation, a fragmentation system was developed for organic chemistry, and a similar approach was applied to other research materials. This has left a vocabulary of about 10,000 terms which, though nearly as specific as before, is far more usable.

We expect terms to be added gradually, but we do not expect to produce a vocabulary in excess of 12,000 terms for several years. The primary consideration in all changes is to increase the usefulness of the thesaurus in providing good answers to users' questions.

Jessica Melton,
Western Reserve University

Development of an Educational Thesaurus at Western Reserve University

A thesaurus of the terminology of education is in preparation by the Documentation Center of Western Reserve University under

the supervision of Gordon Barhydt and Alan M. Rees. The project is the result of a contract with the U. S. Office of Education. The thesaurus will encompass all aspects and subject areas of education and will be used in conjunction with the proposed coordinate indexing system of the USOE's Educational Research Information Center (ERIC). The arrangement of terms will follow faceted lines to provide the basis for the future development of a faceted classification.

It is anticipated that construction will follow much the same pattern established by the Engineers Joint Council in preparing its thesaurus of engineering terms; subject experts and specialist committees will be used to provide validated terminology. The present schedule calls for a working thesaurus by June 1966 which might serve as a basis for development of microthesauri in specialized subject areas.

Josephine L. Walkowicz,
National Bureau of Standards

RICASIP—National Bureau of Standards Information Technology Division

The Information Processing Reference Service of the Research Information Center and Adivsory Service on Information Processing (RICASIP, a joint NBS-NSF venture) is responsible for maintaining and providing access to documentary material on ongoing research and development in the field of information storage, selection, and retrieval. At present, the collection encompasses such areas as documentation, indexing procedures and mechanization, reprography, linguistic analysis, machine translation, pattern recognition, automatic composition, automata theory, and the man-machine interface. Selection of literature for inclusion in the collection is based on subject content rather than on form of publication.

To date, subject access to the collection has been maintained by a very broad classification scheme consisting of 45 categories. The basis of more detailed subject access is being established by

indexing all "current" inputs, i.e., those accessioned since July 1, 1964. As of April 15, 1965, current accessions amounted to 2,700 documents all of which have been indexed in the authors' own words, with no restrictions as to the number of terms assigned to any document. Instructions to analysts specify identification and listing of all important concepts--the criterion of "importance" being whether the document in question would be desirable if retrieved in a search under the term assigned.

In its present form the vocabulary consists of approximately 5,000 terms that now await editing and consolidation into a thesaurus which will serve as an indexing aid as well as a search tool. Computer programs are being written for machine editing of the terms once the vocabulary structure shall have been defined.

PANEL SUMMARY

Joint Panel Discussion

Y. S. Touloukian
Purdue University

Part I: Input Processing

The discussion of input processing could be considered somewhat slanted towards large system operations because most of the panel members were designers or operators of rather large governmental or industrial systems. The summary agenda which was considered during a total of five hours of discussions included these topics:

- Input Media
  comparison of the relative merits of paper tape, magnetic tape, and punched card inputs.

- Storage Media
  criteria for selection of magnetic tape, disc, drum, photo memories, magnetic core, and magnetic card as storage media.

- Character Reader
  practicality of character readers to process inputs.

- File structure
  file organization as a function of system efficiency, with emphasis on file structure, list processing languages and random access storage devices.

Input Media

Most of the discussion of input media centered on punched card and paper tape media. It was agreed that selection of media must be made in terms of the specific characteristics of a given application, and that no universal superiority of one media over

the other existed. But there was a feeling among the group that punched cards, on the one hand, preserve certain advantages of flexibility, reliability, and ease of verification, while magnetic tape, on the other hand, possesses the advantage of speed.

Newer media were also briefly discussed; namely, print reader, Stenowriter and speech transcriber. The print reader, or character reader, was felt to be sufficiently well developed and reliable for use with controlled fonts and formatted material; however, its excessive cost does not make it a widely acceptable input medium at this time. It was felt that the Stenowriter could have specific applications but was still several years in the future. Tape production from the Stenowriter, using an optical scanner, was also anticipated. No views could be recorded concerning audio input media. There was some feeling that mixed systems (e.g., card-tape combinations) could prove attractive for certain applications. The quality and reliability of tape (magnetic and paper) received some attention, and the panel concluded that although high-reliability paper tapes were available, they should be chosen with some care as to their density. No factual data were known to the group on other safety features of tapes.

Storage Media

The group did not find it feasible to compare or to set comparison criteria for magnetic core, card, tape, disc and drum storage media. Such independent criteria as rapid access vs. high reliability vs. cost vs. capacity seem to dictate the respective selection of storage media for specific applications. While it was agreed that magnetic tape did lead the list in having the most advantages, with the exception of rapid access, the group felt there was sufficiently frequent justification to select addressable storage media, such as disc or drum, rather than tape storage. There seemed to be a feeling that newer systems, such as the "Data Cell" and ITEK's photographic memory system, "Memory Centered Process," were rapidly coming into the picture for serious contention as storage media.

File Structure

This phase of the discussion led directly into the request processing aspects of the group's deliberations. Indeed, input and re-

quest processing could not be logically separated. Hence, the section on Request Processing covers the sense of these deliberations in somewhat greater detail. It suffices to state here that seldom, if ever, is a file used for a single-purpose operation. Because of the multiple-purpose use of a file and associated search patterns, its organization is often a compromise arrangement. It was recognized that the structuring philosophy differs when one considers master vs. satellite files, and that the degree of structuring does affect file maintenance and is dependent on the storage media used (i.e., core or tape).

Because of the very nature of the deliberations, no specific conclusions and recommendations were made on the subject of input processing.

William A. Barden
Defense Documentation Center

Part II: Request Processing

The initial topic of consideration in this discussion was that of query phrasing. This was followed by discussions of search logic, security requirements in searching, textual searches, and search output.

Query Phrasing

Not surprisingly, the first point made in the discussion of query phrasing was, "Does the user know what question to ask the system?" As most information system operators will readily agree, this question has at least two meanings: First does the user know *what* he wants? Second, does he know *how* to put the question to the system?

We recognize that many self-appointed critics of information systems claim that a system should be so simple that any user can query it directly. Without going into a long discussion of why's and wherefore's, I can state that it was the consensus of our group that this is not practicable. In our experience an intermediary is required, for two reasons: the first is to help the user specify what he wants; the second is to state the requirement in the language of

the system and in accord with the conventions of the system. This is the most common way of providing for proper phrasing of a query. Normally, the intermediary has a good technical understanding of the subject matter involved and is thoroughly familiar with pertinent aspects of the system. Ideally, he also has experience in the analysis of indexing of the material presented. Thus, he not only understands how things are *supposed* to be done; he also has a good appreciation of how things *are* done in practice.

There is a type of system in which the objective is to permit direct access by the user. This is the "command and control" system, which in our terms of reference represents a highly specialized type of system. The sophistication required is much greater than that found in the real world of information or document retrieval systems. Correspondingly, the cost is a few orders of magnitude greater for both software and hardware. The intermediary is eliminated in this type of system. However, the functions he performs are carried out by the user and the hardware by means of a dialogue, which basically has to be programmed into the system so as to lead the user to ask the right question in the right way.

To get back to our real world of information and document retrieval, the intermediary performs his functions in three ways. First, he helps the user *determine the content* of the query by guiding him on the basis of his own knowledge of the collection and his understanding of the user's stated request. Second, through his knowledge of the vocabulary of the system, the intermediary *expresses the user's requirement* in the terms of the system's vocabulary. Third, he *structures the query* in a form which, when converted into machineable input, will result in a search or series of searches which may satisfy the user. If the results are not satisfactory, it may be necessary to reanalyze the user's stated requirement, reformulate and rerun the search.

In the case of a rather large collection of material in a subject field which has been indexed in depth, it may be useful to specify links and roles in order to limit the results of a search. It would be presumptuous to attempt to cover the subject of links and roles in this paper. Suffice it to say that the use of links and meaningful roles, such as "chemical roles," may justify the increased costs of indexing, input processing, storage, and retrieval by virtue

of improved relevance in search results. However, it is debatable whether either or both will result in improved recall.

Search Logic

The next major element to be considered was search logic. In this area the subject of multiple hits was brought under consideration. In a system such as that designed by SDC,[1] the practice is to run most searches by using only the "or" relationship among many terms. In this approach the output can be ranked by weighting the terms assigned for retrieval. Items identified by multiple hits might generally be considered to have greater pertinence to the request than those items identified by fewer or even single hits. However, the assignment of weights to one or more terms is very effective in the ranking process. Ranking of output is very desirable when the user population consists of people who cannot or will not take time to review large numbers of documents.

In the particular system designed by SDC a number of "must" terms may be specified by the requester. This has the effect of specifying the "and" relationship among those terms. This, of course, is very effective in limiting the results of the search. However, if a request specifies only one "must" term, then a hit based only on that term is likely to be as important as hits on another document which may be identified by other specified terms plus the "must" term.

In other systems, the "and" relationship is most commonly used to specify levels of coordination. Even in these, the "or" relationship is used to represent a concept that cannot be expressed precisely in the vocabulary of the system. Thus, these systems tend to use the "or" relationship within a level of search while establishing the different levels by means of the "and." Practically all mechanized systems provide for negation, i.e., "and not." However, caution is necessary because negation can materially reduce recall. If the system does not use links, the negation is especially risky, simply because the negated term(s) may have no relationship to the desired terms in a particular document.

Again, some systems weight terms assigned to the documents at the time of indexing. DDC does this and finds it can be very effec-

[1] The reference here is to CIRC, p. 15.—Ed.

tive in limiting search output. Here, too, the bibliographers shy away from specifying one or more weighted terms for a search because they know this will reduce recall.

Security Requirements in Searching

Finally, in the area of search logic some attention was given to the matter of the requester's security clearance and need-to-know in relation to the security classification and release criteria of the material being searched. This imposes a requirement in DDC's system and, we are sure, in other systems for using the computer to provide the title sheet for machine-produced bibliographies, to print the highest classification of material in the bibliography, and to produce the necessary receipt forms for classified material.

Textual Searches

Our next topic of discussion was textual searches. Any successful application in this area implies two basic requirements. The first is material available in machineable form. The second is a satisfactory approach to handling the problems of semantics, meaning, and all that these terms imply. Most of us have heard of research and experimentation in the field of textual searching. SDC is developing a system for the Los Angeles Police Department which will use teletype reports of robberies and other crimes. This provides the machineable input without imposing any additional workload. However, on the other side of the coin, the ability to handle the semantics satisfactorily is the area in which SDC is putting forth a considerable effort. SDC feels that progress has been made and is confident that the problem can be solved. We wish them the best of success and will be looking forward to reports on this very interesting project.

The University of Pittsburgh has a project in the retrieval of information in the field of law. I am tempted to call this "legal" retrieval, but this would imply that the rest of us are engaged in "illegal" retrieval. And, of course, this also suggests the possibility that someone might start a system for retrieval of theatrical information. In such a case, would that portion dealing with the stage be "legitimate" retrieval? If so, would the rest of us be guilty of "illegitimate" retrieval? Getting back to the serious aspects of our subject—it is quite possible that exploitation of the association

factor techniques being developed by Ed Stiles[2] and others may be potent for handling the problems of textual searches.
problems of textual searches.

The prime advantage of textual, or natural-language, searching is that it eliminates the need for (1) a thesaurus, (2) analysis of material during input, and (3) an intermediary. But the price of attaining these advantages is likely to be high, since very large-scale data processing hardware and very exotic programming are required to cope with this approach. It may be too costly for general use in information and documentation retrieval systems.

Search Output

Finally, we turned our attention to the subject of search output. In terms of the results of individual searches, the output can take any of several forms. For example, it could be a listing of the descriptive information or any designated elements of this information for each identified document. It could list only the indexing terms plus the accession number. It could list only the abstract plus accession number. It could list any combination of these.

On the other hand, the output could be in the form of punched paper tape which carries the identification of the requester, the search number, and the accession numbers of documents identified by the search. This tape can be put on Telex or TWX machines for transmission to distant points, and the user can look up the entry for each accession number. DDC is using this on a number of searches each day, and the results are transmitted to the field office from which the request was received. We call this service "Telex Searches."

Another means of providing search output is by a device such as the SC 4020. This is less expensive than the conventional mechanical printer and is of particular interest, in terms of requirements, for obtaining multiple copies of the output. Like film processing, it has the disadvantage of lengthening the response time of the system. I doubt though that this would be critical.

Another method is photocomposition. DDC and the Clearinghouse of the Department of Commerce are preparing to go to photocomposition of announcements about July 1965. It is quite prob-

---

[2] Mr. Stiles was a participant from the Department of Defense. —Ed.

able that both organizations will be using this method to produce copy for printed bibliographies by December.

All of these methods except punched paper tape output involve the use of a computer as a rather expensive printing device. In the case of direct output via mechanical printers the computer is not very satisfactory printing device.

We believe that ultimately we can have semiautomatic photographic storage and retrieval, at least of the entries, and when this time comes we will terminate the computer search by punching a card which identifies the requester, the search number, and the accession number of each document identified by the search. A deck of these cards will then be placed in the hopper of the semiautomatic photographic system to trigger the production of photographic or electrostatic enlargements of the bibliographic information for each entry. The semiautomatic photographic storage may turn out to be a video file of the necessary images. At present, however, video-file techniques leave much to be desired in the readability of text.

Another form of output involves producing a deck of punched cards and manually pulling the corresponding catalog cards to complete the request. The problems involved in stockpiling catalog cards, that is, in reproducing the cards and maintaining accountability for the classified cards, are not to be considered lightly. We in DDC gave the matter serious consideration and put an end to that particular bucket of worms. Most of our customers were delighted. To the best of my knowledge, none of them has protested the move.

I wish to thank all the participants in this group for their valuable contributions to the discussions which made this summary possible.

PANEL SUMMARY

Van A. Wente
National Aeronautics and Space Administration

Announcement and Dissemination

The panel members interpreted *announcement* to mean prompt service in bringing new documents to the attention of probable users by whatever means possible, including manual as well as semiautomatic means. They treated *dissemination* similarly by concentrating on the best means of providing users with whatever form of information was desired: a bibliographic description, an abstract, or a full text.

Announcement

In discussing announcement service, the panel listed the following six commonly used media and noted the advantages and limitations of each:

- Personal service by an information specialist.
- Computer-prepared notifications based on user profiles.
- Standard accessions lists not tailored to any particular user.
- Awareness-provoking announcements, such as an in-house newsletter covering documents, projects, personnel.
- Topical abstract-index journals, including both discipline-oriented, as *Chemical Abstracts* and *Physics Abstracts,* and function-oriented, as a journal for a specific industry, such as the tobacco industry.
- Special-purpose abstract-index journals, such as *Nuclear Science Abstracts, Technical Abstract Bulletin* and *Scientific and Technical Aerospace Reports.*

It was agreed that the quality of each service varies widely. Personal service by a professional information specialist, for example, ranges from the best possible quality to the worst possible. Computer-prepared notifications also vary in quality, but not quite so much. Least variable for all media are the abstract-index journals.

The panel observed that the quality of computer-prepared notifications is affected by several factors. These include (a) the no-

tification format, usually a card or a list; (b) the amount of information provided, which might include, for example, only a bibliographic description or an abstract; (c) the ease of obtaining a requested full text; and (d) the quality of matching a user's profile with the indexing terms of a document. This quality factor is affected by the size of user and document populations and by the methods used to formulate and maintain user profiles. Profile quality is of vital importance.

Computer-prepared notification service is not presently provided as an across-the-board service to all users. Most users by necessity receive incomplete service. But the "elite" scientific and engineering users should, however, receive this kind of individual service. The panel agreed that, ideally, automatic profile matching would eliminate the need for all other announcement services to the "elite" users.

The panel agreed that profile generation should always begin at the level of the individual user. If profiles are grouped for economy, the homogeneity of the selected users must be respected. Controlling principles for valid grouping are (a) organizational unity, as found among users in a laboratory; (b) group characteristics of the users' professional discipline(s); and (c) common mission-orientation of the users. A point underscored by the panel was that consolidation is an irreversible process; there is no way to reconstruct individual profiles once they are merged according to a common principle.

The panel further agreed that the service of matching profiles with document index terms should not be regarded as a mechanism for absorbing costs of indexing or input, which have to be done anyway for other information services.

The abstract-index journals, as previously noted, vary less in quality than do other announcement media. This is due in part to the type of users they serve: Type A, the energetic reader, eager to scan many titles, abstracts, or indexes; and Type B, the user who wants only a small selection of information.

The panel reached several conclusions on abstract-index journals, all of which can be summed up by the general statement that this announcement media is very difficult to handle. With respect to categories used for the grouping of information, the panel agreed

that these journals cannot be properly located by a single category and that category assignments should be supplemented by multiple listing or by easy-to-use indexes. Assignment of subject categories generally implies an impossible degree of foresight and at best is a temporary measure. Multi-discipline journals should be published in separate category groupings, or in several separate categories if abstracts are multiple-listed.

Regarding indexes, the panel felt that cumulations of indexes are needed by both the information specialist and the scientist/engineer user and should be tailored accordingly. Printing quality and ease of reading are important but secondary to the content of the information being presented. Once an economically feasible level of printing quality has been selected, this level should be changed only upon significant complaint of the users.

Premuted title indexes are useful primarily for temporary or special purposes, but any enrichment or addition of terms to permuted title indexes is tantamount to normal indexing and almost equal in cost. The panel therefore questioned the use of permuted indexes in place of normal indexes.

Dissemination

Dissemination of information was divided into three types, at the levels of (1) bibliographic data, (2) the abstract, (3) and the full text. It was noted that any announcement-dissemination service can expand its scope incrementally by adding each level until all three are included. General conclusions of the panel covered microfiche forms, easier ordering service for users, and improved dissemination service.

Microfiche, at a cost of less than 10 cents, will have a great impact on full-text dissemination to organizations. And if costs are reduced by mass production or by new techniques, such as off-set printing, microfiche might even be used in computer-determined *individual* dissemination.

The panel recommended an alternate, or perhaps parallel, dissemination scheme that might prove useful to users who want a complete document. It would work this way: At the initial printing of a new report, a postcard order form would be mailed to individuals or groups possible having interest in the report. The number of returned postcards would indicate the degree of interest to the

originator of the report. He could then better estimate the quantity to be printed both for dissemination to persons on contract-specified lists or standard distribution lists and also to postcard requesters. Extra copies of the document could be stocked for postcard requests sent in at a later time.

The panel was of the opinion that over-all dissemination of documents could be improved, with resultant savings, through the cooperative exchange of machine-readable document descriptions by information producers and sponsors. In addition, even greater improvement and savings might be realized if announcing services exchanged individual user profiles for computer-prepared notifications and announcement. Thus, as a final conclusion, the panel recommended exploring the feasibility of establishing central files of machine-readable user profiles.

PANEL SUMMARY

Fred H. Wise
System Development Corporation

User-System Relationships

When a document store or data store becomes so large or so complex that the information seeker can no longer efficiently use it, new relationships between the user and the information center inevitably develop. The purpose of this panel was to exchange ideas regarding the nature of these relationships. Participants represented Government, industry, non-profits, and the academic world; the represented information systems ranged from a highly specialized science and technology store to an information system designed to enhance public relations. The discussions can be summarized under three main topics: (1) development of an atmosphere conducive to user confidence in, and utilization of, the system; (2) effective communication between the user and the system; and (3) productive feedback for system evaluation and improvement.

User confidence

With certain notable exceptions, there was agreement among the group that there is a general lack of confidence in technical libraries and information systems. If this situation is to be changed, information systems must actively provide services otherwise unavailable, and not passively serve as repositories of information. An information system cannot be forced upon the user; his confidence must be developed before he will accept it.

The more a potential user is involved while a system is developing the stronger will be his interest and the greater will be his confidence. Often, the lead time is great and it is difficult to arouse the interest of the potential user. He should, therefore, be involved in preoperational phases to assure that he will patronize the system after it is in operation. The user frequently does not have a clear picture of his needs, because he is deeply influenced by the system he is currently using. However, he can best identify his require-

ments empirically; that is, through actual use of a pilot system. The panel recommended that despite these and other difficulties, every possible effort be made by the system designer to interact with the potential user. The designer should be able to obtain, at least in general terms, the types or categories of required information, the size and complexity of search or reference questions, the number and frequency of searches; and he should also find typical examples of successes and failures of the current system.

Designing a system for various levels of users is another consideration in developing an atmosphere to induce users to patronize a system. The new user will may have different requirements than the experienced user has. The user on a managerial level may have still another requirement. It was suggested by the group that flexibility in design is essential to meet a variety of requirements. The user must be carefully identified and his requirements analyzed. Too often this has not been done.

Another factor in the motivation of the user to utilize the system is the degree to which he understands its scope. Scope may best be defined in terms of the content or input of the system, the identification of the users it is designed to serve, and the types of service it provides. Provision should be made for orientation and training of the user. He should be given progress briefings prior to the operational date. Orientation sessions should include demonstrations of the system's capabilities and services to be provided. Where possible, personal contact between system personnel and the user should be encouraged as a means of training and orientation.

It becomes clear that in large systems, whether manual or semi-automated, others must perform services for the information user. For example, the user may need to delegate such functions as preliminary screening of input, searching and retrieval, and selection of pertinent documents from the retrieval output listing. When functions are delegated, a host of problems relating to motivation become crucial. Professionally trained scientists and engineers, when forced to depend upon unqualified personnel for assistance in conducting their searches, develop unfavorable reactions to the system.

The panel participants generally agreed that the value of the well-qualified information specialist should not be underestimated.

They agreed that there is a high investment in these people. The role of the specialist varies with the type of information being handled by a specific system, and training requirements vary widely. There is a strong need to define the information specialist as a professional and to take all possible steps to enhance his status. Training programs should be developed, appropriate college courses should be encouraged, and corresponding high school indoctrination should be conducted to attract high-caliber personnel. Further, practicing information specialists should be encouraged to develop their own stature through the production of quality literature reviews.

An entirely different training problem is encountered when the output of one system becomes the input of another system. Discussions of this special case indicated it is expected to become increasingly important as more systems interface with one another.

Communication between User and System

Communication between the user and the system may be either direct or indirect–direct, in the sense of the user utilizing consoles for input and output; indirect, in the sense of his using the services of an information specialist to do the communicating.

It was generally agreed by the participants that direct communication is difficult to manage and is often too expensive for general use. However, with small or homogeneous user groups it can be very effective. In this case, it was recommended that designers exploit telecommunication devices in the field of information science. Some systems of note are now using such devices; for example, the National Lending Library in England, the Los Angeles Police Department, and the U. S. Naval Ordnance Test Station in China Lake, California, which is carrying out pilot studies. Time-sharing offers some interesting possibilities for expanding direct communication to a much broader spectrum of users.

Indirect communication presents another set of problems. Rapport between the user and the information specialist is very important. The user should be free as possible to select the information specialist who best meets his needs. The user must have confidence in the abilities of the information specialist.

The nature and format of the outputs from an automated system

also are important factors in indirect communication. The outputs must be designed for optimum usability. They should be in natural language, have good quality print, be convenient to handle, and have the data ordered for optimum use. Further, a variety of output formats is generally required when a large population of users is served. Flexibility in output is basic in meeting the requirements of most user populations, and the initial cost of design and implementation of a flexible system is more than repaid in user acceptance and utilization.

System documentation, another form of communication essential to the life of a system, was discussed by the participants. Complete system documentation is necessary so that the loss of key personnel will not necessarily spell collapse of the system. Furthermore, modifications are more readily accomplished when complete documentation is available. In view of this, it was recommended that buyers and contractors take into consideration the time and cost of documentation from the very beginning of the design effort. Also, system operators and maintenance personnel must be made aware of the need to document changes or modifications thoroughly, regardless of their scope.

Feedback for Evaluation and Modification

Feedback is essential for system evaluation and modification. Too often, well-planned and well-organized feedback is downgraded in the design of information systems. The group agreed that no one system of feedback will be productive for all systems. A definite plan for the collection, analysis, and utilization of feedback data is vital to a dynamic system if it is to be sensitive to the changing needs of its user population.

The nature of the feedback is dependent on the specific system. It was generally agreed that lengthy or time-consuming questionnaires generally defeat their own purpose. Personal contact between the user and operations personnel is highly recommended as a feedback vehicle. When carefully planned, and when rapid analysis is possible, a short checklist has been found to be effective as a feedback vehicle.

Motivating the user to provide feedback presents a major problem. It was generally agreed that the user wants to see the results

of his feedback reflected in changes in system effectiveness. The user wants to know what action has been taken on his feedback, whether it is negative or positive. It was therefore recommended that the user be provided feedback on *his* feedback at every opportunity.

PANEL SUMMARY

James W. Singleton
System Development Corporation

System Parameters and Management

The modus operandi of the panel was to list critical development milestones and to discuss each in terms of the experience of the participants. This resulted in the identification of several problems, and agreement on two recommendations.

Education of Management

A common problem appears to be lack of support and understanding on the part of management regarding development of an information storage and retrieval system. This may be due ir part to the fact that information services are traditionally and characteristically inexpensive. It also may be due in part to lack of information about the types and extent of information service which a modernized information storage and retrieval system can offer. In any event, the panel agreed that management is acutely cost conscious and does not encourage innovation. It was suggested that professionals in the field have not paid enough attention to the marketing or public relations requirements which are necessary to alter this set of management attitudes.

System Evaluation

Animated discussion was held on the problem of evaluating the effectiveness of an information storage and retrieval system. Opinion ranged from the view that subjective evaluation is adequate ("If the user likes the system, it is a good one") to the view that quantitative measures, such as the proportion of dissemination "hits," are the only index of system effectiveness. Most participants would not exclude the user's satisfaction as a criterion in system evaluation, but they generally agreed that lack of adequate statistics in this matter poses a problem and that more quantitative measures are required. (See recommendation 1.)

Design Control

Discussion of this topic centered on the question of the basic objectives for which information storage and retrieval systems are developed. Concern was expressed that, too often, a system is designed by documentalists simply to mechanize the traditional processes of information services. By contrast, more attention to user interests will usually result in system design that is innovative and non-traditional.

What is the problem?

A pitfall that several participants had experienced is that of attempting to solve the wrong problems with a mechanized or automated information storage and retrieval system. Management and organizational problems, such as those affecting lines of reporting and scope of delegated authority, will not be solved with an information system, and may even be aggravated. This can result in a black eye to the information system even though it might actually be a technically competent system.

Costs of Automation

It is sometimes argued that automation will produce savings in performing information services. Such savings are rarely realized. Typically, an improved information system results in the availability of new capabilities and services which had not been performed previously and which eat up potential savings. Such new services are often highly desirable. Their implementation and use, however, depends on a design decision that properly rests with the project management who should clearly recognize the cost.

Personnel Requirements

Personnel to operate and maintain a mechanized or automated information system require more specialized training than did their predecessors in a manual system. This, together with the growth in numbers of automated systems, has resulted in a severe shortage of competent personnel. One panel member reported good success in cross-training of engineers and subject matter specialists. Other members stressed the necessity for investing the career fields of information services, such as cataloging and indexing, with greater prestige and importance. This should be accompanied by continu-

ing emphasis on creating career opportunities and salary advancement within these fields. One person reported that he had had a new Civil Service classification recently approved for this career field.

Operator Training

Closely related to the preceding point is the requirement to train information system operators, during system development, in the operation of a new system. This should be initiated well before turnover of any major system feature to assure continuity of day-to-day performance when installation is completed. Provision for such training should be contained in the early system design to assure that it is not overlooked or viewed as an afterthought.

System Transition

Differences of opinion were clearly evident among panel members on the question of how to accomplish cut-over, or transition, from a manual system to an automated information system. On the one hand, it was recognized that an organization manager does not want to risk interruption of current operations by disbanding the predecessor system prematurely; on the other hand, extended parallel operations were viewed as expensive and hazardous because they remove the incentive to *make* the new system work. No completely satisfactory solution was identified by the panel, although agreement was noted on the necessity for well-planned and thorough quality-control testing procedures as an inducement to more immediate cut-over scheduling.

Breaking the Caste System

A common problem in the development of a mechanized or automated information system is that of user involvement—how much, when, in what role. Concensus of the panel was that it is usually difficult to get the user involved in any significant capacity and that, therefore, a severe communication barrier is created that results in a usable system going unused. Several of the panel members reported good results from a campaign of public relations that kept the information system visibly (and favorably) in front of organization management. Specifically, one information services director stated that he never lets six months pass without implementing a management-oriented information system feature.

Achieving an Integrated System

Anomalies occur often in the early operation of an automated information system. One panel member reported that first reaction on the part of technical personnel was favorable when they received bibliographies and accessions lists prepared by machine. Upon going to document control to order a periodical, however, requesters were asked to fill out traditional request forms that completely failed to take advantage of machine retrieval capability and, further, had to be prepared in triplicate for each accessions number. This quickly produced disenchantment. The point is that a narrow view of the scope of an information storage and retrieval system can result in the overlooking of such important considerations as methods and procedures for ease of customer use. A more desirable approach is to take a realistically broader, user-oriented view in order to assure uniform, integrated system operation.

System Design Optimization

General agreement was reached among panel members that alternative design objectives of information systems are to optimize function, cost, or speed. Some systems deliberately attempt such an optimization, as in the case of some military systems wishing to achieve the greatest functional capability within the state of the art. Most systems are, of course, a compromise among these alternative design objectives. The point which the panel wished to highlight is that these alternative design objectives are competitive; emphasis on one results in trade-off with the others. Since all cannot be optimized, management decision must be the basis for selection among them and must be clearly established in setting forth design goals.

ADP Management and Reporting

It was clearly recognized that trends of the past several years are toward multiple-use computer installations. That is, information storage and retrieval systems are being designed for document indexing and dissemination, accounting and inventory control, management information reporting, research and scientific computation, etc., all within the same system. This poses the organizational and administrative problem of managing a computer center such that all users receive equal attention and are charged fairly for their

share of the costs of operation. While this is not an insoluble organizational and administrative problem, it is a complicated one. Very little empirical experience is available on which to base decisions.

Establishing ADP System Requirements

Establishing requirements for an ADP system is a classical problem beset with uncertainties regarding validity, detail, and scope. The panel had no unique solution to offer. On one point, however, quite marked agreement existed which should be noted. At least for the class of ADP systems that perform library and information services, it was believed that user studies and survey do not constitute an acceptable criterion for design. The field is so unstructured and so many different sets of users exist that it is possible to prove almost anything by a user survey. User studies were recommended by the panel as being of primary value in the public relations accompanying a system development in the important effort to dispose users and management favorable toward a system.

Peripheral Payoff

As indicated by their comments on achieving an integrated system, the panel did not take a restrictive view of the scope of an information storage and retrieval system. One important way this was expressed was with respect to the peripheral operations of n information system, e.g., receipting, security control, mailing and records keeping. These operations can often be mechanized or automated in the course of system development, at little or no incremental cost, with great improvement resulting in ease of operations and high visibility to management.

The Problem of Purging

As high a degree of interest was generated by the panel on this issue as on any other topic of the workshop. It was said that rules for record and document retention usually don't exist and aren't followed if they do. One participant asserted that he would recommend, from his experience, that no purging be performed, but that retrieval response rates simply be extended for older and less-frequently used materials. (See recommendation 2.)

Backlog Data Conversion

Several panel participants could speak to the topic of backlog data conversion with the voice of experience. Each of these shared the same opinion: "Don't do it." They said that an analysis of the requirements should be performed only for special cases (e.g., selected categories of classified data), and expressed the view that it is far better to start a mechanized or automated system with a clean slate, incorporating only current data and leaving requests for older data to be filled by manual search techniques.

Recommendation 1: System Evaluation

The panel recommended development of a technology for evaluating the total impact of automatic data processing in a specific information storage and retrieval system. This technology should be based on *system* performances measures, and not on machine or computer program capability alone.

Recommendation 2: Purging of Files

The panel also recommended development of a general theory of file purging with which techniques of implementation can be associated. Such a theory should differentiate the character of the data base (as, for example, classified, archival, or legal) and retention rules based on such factors as time of source document publication, frequency of use, reliability of source, and retrieval response requirements.

APPENDIX

Appendix I

Biographies

William Barden is Special Assistant for Operations Research to the Administrator of the Defense Documentation Center. Mr. Barden has held several positions with DDC and its predecessor agencies and was one of the original members of the group in London at the end of World War II which processed captured German technical documents. This effort evolved into the present organization known as the Defense Documentation Center.

Raymond P. Barrett, Manager of the Technical Information Systems Department, System Development Corporation, is responsible for SDC's contractural operations involving the national intelligence community. He is also responsible for activities in library and documentation systems, natural language data systems, and information systems supporting science and technology Mr. Barrett has had more than fourteen years in the intelligence field, primarily with Government agencies.

Bernard K. Dennis is affiliated with the Battelle Memorial Institute in Washington, D. C. He is in charge of the Engineers Joint Council course of instruction on "System of Roles for Information Retrieval." He was formerly Manager of the Technical Information Center, Flight Propulsion Division, General Electric Company. Mr. Dennis managed one of the first operational information storage and retrieval computer systems.

William Hammond is Manager of Computer Operations at Datatrol Corporation. Formerly he was with ASTIA (now DDC) as chief of the Directorate of Automatic Systems and Services. Mr. Hammond helped establish DDC's computer system.

William T. Knox is Technical Assistant to the Director, White House Office of Science and Technology. Mr. Knox has previously been a consultant on technical information matters to the National Science Foundation, Engineers Joint Council, and the Department of Defense. He holds several patents and

is the author of several articles on petroleum engineering and information research and management.

C. Allen Merritt is Manager of Selective Dissemination and Microprocessing at the Thomas J. Watson Research Center of International Business Machines Corporation. Mr. Merritt has been affiliated with IBM since 1956 when he joined the Corporation as manager of the Publications and Information Department in Poughkeepsie, New York.

Herbert Rehbock is Director of Document Analysis and Processing at the Defense Documentation Center. He has been with DDC and its predecessor agencies from the time the first agency was established at Wright Patterson AFB. Mr. Rehbock has held various executive positions within DDC and is an acknowledged expert on document processing.

James W. Singleton, System Development Corporation, is responsible for several contracts with military and civilian agencies of the Government. Dr. Singleton has worked on the development of the SAGE air Defense System and NORAD Combat Operations Center, and has assisted the Air Force in conceptual analysis and planning studies leading to the next generation of command-control systems. Dr. Singleton has written and lectured on the subjects of command-control, simulation of man-machine systems, system training, and the management of computerbased information system development projects. He is a member of the USAF scientific advisory board.

Y. S. Touloukian is Director of the Thermophysical Properties Research Center, School of Mechanical Engineering, Purdue University. Dr. Touloukian has been instrumental in developing methods and techniques for the searching, coding, and mechanical processing of bibliographical information on the thermophysical properties of chemical compounds.

Van A. Wente, a member of the Documentation Branch, Scientific and Technical Information Division, National Aeronautics and Space Administration, is concerned with the administration and development of NASA's information processing activities. Mr. Wente has worked for the Firestone Tire Company and the Naval Research Laboratory in research and

development activities. He also has worked with the Atomic Energy Commission in the field of information science.

Audrey S. Williams is Assistant Supervisor of the Development Engineering Subdivision library, Missiles and Space Division, Douglas Aircraft Company. Mrs. Williams has been with Douglas Aircraft since 1960 and is concerned with the Douglas automated documentation system.

Fred Wise, System Development Corporation, is responsible for the design, development, and implementation of a personnel subsystem for an information storage and retrieval system. He is also responsible for the development of the scientific and technical vocabulary by which a user can communicate with the system. Mr. Wise has had broad experience in analysis, design, and implementation of several computer based systems.

Harold Wooster heads the Information Services Directorate, Air Force Office of Scientific Research. Since 1956 Dr. Wooster has been concerned with selection and support of long-range programs of basic research in the information sciences.

APPENDIX

Appendix II

Workshop Participants

Lida L. Allen
Irving Aronowitz
Manuel Avila
Capt. Charles W. Back
William A. Barden
Glen R. Barnard
Lester A. Barrer
Raymond P. Barrett
G. W. Beveridge
L. M. Bohnert
James J. Brady
Capt. T. K. Burgess
Howard Burnaugh
Roger VanBuskirk
W. M. Carlson
W. D. Climenson
Al DeLucia
B. K. Dennis
John A. Dovel, Jr.
Sheldon G. Ericksen
George Ember
James H. Fisher
Earl G. Fossum
Roger E. Graves
Edward K. Grimes
William Hammond
Leo Harris
Mary Herner
Percy B. Hilburn, Jr.
A. G. Hoshovsky
Robert D. Igou
Paul Irick
G. Jahoda
Alec C. Jones
Leonard Karel
George G. Kershaw
William T. Knox
Mitchell A. Krasny
Frank P. Krasovec
Frank J. Kreysa
Carolyn Kruse
George F. Lewenz
William M. Lyons
LeRoy B. McCabe
Marvin W. McFarland
H. S. McMann
W. McMillan
John L. McNamara
Jessica Melton
C. Allen Merritt
John S. Moats
Daniel D. Moore
Joseph D. Naughton
Joel S. O'Connor
W. G. Patton
Ralph L. Peterson
George R. Pielmeier
Lt. Col. Davis B. Potter
Pauline C. Ramsey
Herbert Rehbock
Robert P. Rich
William H. Richardson
C. David Rife
Robert Root
Terry R. Savage
Capt. Harlan L. Sailor
Charles J. Schmidt
H. R. Seiden

Victor F. Shauklas
John Sherrod
James W. Singleton
Samuel S. Snyder
Madeline S. Startzman
H. Edmund Stiles
Edmund J. Sumpter
R. Swanson
Sarah M. Thomas
Harold B. Thompson
Lt. Col. John A. Thompson
William Thompson

Y. S. Touloukian
Hans C. Ullmann
James G. VanCot
Josephine L. Walkowicz
Richard Watson
Paul A. Wehmeyer
Van A. Wente
Audrey Williams
Fred H. Wise
Harold Wooster
Ronald E. Wyllys
J. L. Zaharias